互いに相関する3変量間の回帰分析について

メタデータ	言語: jpn 出版者: 公開日: 2013-08-27 キーワード (Ja): キーワード (En): 作成者: 野田, 明男 メールアドレス: 所属:
URL	http://hdl.handle.net/10271/2634

互いに相関する3変量間の回帰分析について

野田明男
（総合人間科学講座・数学）

On the Regression Analysis for Mutually Correlated Three Variables

Akio NODA
Integrated Human Sciences · Mathematics

Abstract: Let us consider three random variables X_i ($i = 1,2,3$) with mean μ_i and variance σ_i^2. We denote by ρ_{ij} the correlation coefficient of X_i and X_j, and set $\{i, j, k\} = \{1,2,3\}$. Then the partial correlation coefficient $\rho_{ij\bullet k}$ is defined to be $(\rho_{ij} - \rho_{ik}\rho_{jk}) / \sqrt{1-\rho_{ik}^2}\sqrt{1-\rho_{jk}^2}$ (see [4]), which is equal to the correlation coefficient of residuals $R_{i\bullet k}$ and $R_{j\bullet k}$. Here, we put $R_{i\bullet k} = X_i - \hat{X}_i$, $\hat{X}_i = \mu_i + \sigma_i\rho_{ik}(X_k - \mu_k)/\sigma_k$ being the least squares regression line of X_i given the value X_k. In § 1 we study some properties of these partial correlation coefficients to see their importance in the regression analysis and also in the theory of normal distributions.

The purpose of this paper is to investigate the logarithm v_k of cancer incidence in Japan (due to [2]), which corresponds to the value $x_k = 2.5 + 5(k - 1)$ ($1 \le k \le 18$) of age. The fact that the correlation coefficients between three data $\{x_k, u_k = \log x_k, v_k\}$ are all near to 1 was observed in [1], which surprised the author and led him to the present study of these data. Indeed, we find outliers in the residuals $R_{u\bullet x}, R_{v\bullet x}, R_{v\bullet u}$ and compute the partial correlation coefficients $r_{uv\bullet x}$ and $r_{xv\bullet u}$ to note two remarkable low values: one is $r_{uv\bullet x} = 0.0543$ in the range $1 \le k \le 18$ of age and the other is $r_{xv\bullet u} = 0.2614$ in the range $4 \le k \le 18$ of age, which tells us that our real data v_k can be fitted by the regression line on x_k (resp. u_k) in the former (resp. latter) range of age.

The final section is devoted to a study of the simulated data w_k that we generate by using the Weibull distribution ([3]). Our method of simulation comes from an approximation of the simulation model proposed in [1]. We obtain results on various kinds of partial correlation coefficients such as $r_{uw\bullet x}, r_{vw\bullet x}$ and $r_{xw\bullet u}, r_{vw\bullet u}$ defined by (1–1), and also $r_{vw\bullet xu}$ defined by (1–4).

Key words: partial correlation coefficient, regression analysis, logarithm of cancer incidence, Weibull distribution

§ 1. 偏相関係数，線形回帰の残差分析と正規分布

この節ではまず，互いに相関する3つの確率変数 X_1, X_2, X_3 の間に成り立つ関係式を論じる。X_i の平均，分散を μ_i, σ_i^2 とし，X_i と X_j の相関係数を ρ_{ij} とする（$1 \le i < j \le 3$）。k を $\{i, j, k\} = \{1,2,3\}$ となる番号とするとき，X_k の影響を除いた後の X_i と X_j の相関係数は次式で定義され，偏相関係数と呼ばれる（[4] p.53）。

（1−1）　$$\rho_{ij\bullet k} = \frac{\rho_{ij} - \rho_{ik}\cdot\rho_{jk}}{\sqrt{1-\rho_{ik}^2}\sqrt{1-\rho_{jk}^2}}$$

(X_1, X_2, X_3) が3次元正規分布に従う場合，X_k に基づく X_i の条件つき期待値は，

（1−2）　$$E(X_i|X_k) = \mu_i + \frac{\sigma_i\rho_{ik}}{\sigma_k}(X_k - \mu_k)$$

となり，最小2乗法による線形回帰 $\hat{X}_i = \beta_0 + \beta_1 X_k (\beta_1 = \sigma_i\rho_{ik}/\sigma_k, \beta_0 = \mu_i - \beta_1\mu_k)$ に一致する。このとき残差 $R_{i\bullet k} = X_i - \hat{X}_i$ は，X_k と無相関（正規分布なら独立）であり，その分散は $\sigma_i^2(1-\rho_{ik}^2)$ で与えられる。こうして，X_k の値を知ったとき，残差 $R_{i\bullet k}$ と $R_{j\bullet k}$ の相関係数は，

$$E\Big((X_i - \hat{X}_i)(X_j - \hat{X}_j)\Big)\Big/\sigma_i\sqrt{1-\rho_{ik}^2}\,\sigma_j\sqrt{1-\rho_{jk}^2} = \Big\{\mathrm{Cov}(X_i, X_j) - \beta_1\mathrm{Cov}(X_k, X_j)\Big\}\Big/\sigma_i\sqrt{1-\rho_{ik}^2}\,\sigma_j\sqrt{1-\rho_{jk}^2}$$

と計算できて，上式（1−1）に等しくなることがわかる。

（1−1）を用いると，因果関係で成り立つ推移律を相関関係に対しても議論できる。すなわち，ρ_{ik} と ρ_{jk} が $\min(\rho_{ik}, \rho_{jk}) = 1-\delta$（$\delta > 0$ は小さい）で，ともに1に近いものとする。このとき，$\rho_{ij} \ge 1-\delta$ となるためには偏相関係数 $\rho_{ij\bullet k} \ge \frac{1}{2} - \frac{\delta}{2(2-\delta)}$ が十分条件となる。しかしながら，$\rho_{ij\bullet k}$ が−1に近ければ，ρ_{ij} は $1 - 4\delta + 2\delta^2$ のレベルに低下する。そして常に，$|\rho_{ij\bullet k}| \le 1$ となるので，次の結果を得る。

<u>命題1.</u>　$|\rho_{ij} - \rho_{ik}\cdot\rho_{jk}| \le \sqrt{1-\rho_{ik}^2}\sqrt{1-\rho_{jk}^2}$

次に，偏相関係数は3次元正規分布の密度関数を記述するときに有効であることをみよう。X_i の標準化 $Z_i = (X_i - \mu_i)/\sigma_i$ を用いて密度関数 $f(x_1, x_2, x_3)$ をかくと，$\frac{1}{C}\exp\left[-\frac{1}{2}\sum_{i,j=1}^{3} q_{ij}z_iz_j\right]$ となる。ここで3次の正定値行列 $Q = (q_{ij})$ は，相関行列 $R = (\rho_{ij})$ の逆行列に等しく，正規化定数 C は全範囲における $f(x_1, x_2, x_3)$ の積分が1になるという条件で定まる。$\det R = 1 - (\rho_{12}^2 + \rho_{13}^2 + \rho_{23}^2) + 2\rho_{12}\rho_{13}\rho_{23}$ に留意して q_{ij} を計算すると，$q_{ii} = \frac{1-\rho_{jk}^2}{\det R}, q_{ij} = q_{ji} = \frac{\sqrt{1-\rho_{ik}^2}\sqrt{1-\rho_{jk}^2}}{\det R}\rho_{ij\bullet k}$ を得る。従って，

2次形式 $\sum_{i,j=1}^{3} q_{ij}z_iz_j$ の項は，$w_i = \sqrt{\frac{1-\rho_{jk}^2}{\det R}}z_i$ と変数変換すれば，$\sum_{i=1}^{3} w_i^2 - 2\sum_{1\le i<j\le 3}\rho_{ij\bullet k}w_iw_j$ という3通りの偏相関係数 $\rho_{ij\bullet k}$ を用いた形に書き直すことができる。

（註1）　2次元正規分布の場合は，$w_i = z_i / \sqrt{\det R} (\det R = 1-\rho_{12}^2)$ と変数変換すれば，密度関数の2次形式の項が $w_1^2 - 2\rho_{12}w_1w_2 + w_2^2$ と書けることは見やすい。そして w_2 の値を知ったときの w_1 の条件つき密度関数は，$\dfrac{\exp\left[-\dfrac{1}{2}\left\{(w_1-\rho_{12}w_2)^2+(1-\rho_{12}^2)w_2^2\right\}\right]}{\int (\textit{the same expression as above}) dw_1} = \dfrac{1}{\sqrt{2\pi}}\exp\left[-\dfrac{1}{2}(w_1-\rho_{12}w_2)^2\right]$ と計算され，

平均 $\rho_{12}w_2$ をもつことが従う。

3次元から4次元の正規分布に移る。$\{i,j,k,\ell\} = \{1,2,3,4\}$ とする。2変量 (X_k, X_ℓ) に基づく X_i の条件つき期待値は，偏相関係数(1−1)によって

(1−3)　$E(X_i | X_k, X_\ell) = \mu_i + \dfrac{\sigma_i \rho_{ik\bullet\ell}}{\sigma_k}\sqrt{\dfrac{1-\rho_{i\ell}^2}{1-\rho_{k\ell}^2}}(X_k-\mu_k) + \dfrac{\sigma_i\rho_{i\ell\bullet k}}{\sigma_\ell}\sqrt{\dfrac{1-\rho_{ik}^2}{1-\rho_{k\ell}^2}}(X_\ell-\mu_\ell)$

と計算される。X_i から Z_i，さらに W_i へという上記の変換を行えば，この式は

$E(W_i | W_k, W_\ell) = \rho_{ik\bullet\ell}W_k + \rho_{i\ell\bullet k}W_\ell$ と書き直される。

w_2, w_3 の値を知ったときの w_1 の条件つき密度関数は，$\dfrac{1}{\sqrt{2\pi}}\exp\left[-\dfrac{1}{2}(w_1-\rho_{12\bullet 3}w_2-\rho_{13\bullet 2}w_3)^2\right]$ と計算され，その平均は $\rho_{12\bullet 3}w_2 - \rho_{13\bullet 2}w_3$ に等しい。

さて，(1−3)の右辺を線形2重回帰の式

$$\hat{X}_i = \beta_0 + \beta_{1k}X_k + \beta_{1\ell}X_\ell \quad \left(\beta_{1k} = \rho_{ik\bullet\ell}\frac{\sigma_i}{\sigma_k}\sqrt{\frac{1-\rho_{i\ell}^2}{1-\rho_{k\ell}^2}}, \beta_{1\ell} = \rho_{i\ell\bullet k}\frac{\sigma_i}{\sigma_\ell}\sqrt{\frac{1-\rho_{ik}^2}{1-\rho_{k\ell}^2}}, \beta_0 = \mu_i - (\beta_{1k}\mu_k + \beta_{1\ell}\mu_\ell)\right)$$

に転用して，残差 $R_{i\bullet k\ell} = X_i - \hat{X}_i$ を求めると，その分散は $\dfrac{\sigma_i^2}{1-\rho_{k\ell}^2}\left(1-(\rho_{ik}^2+\rho_{i\ell}^2+\rho_{k\ell}^2)+2\rho_{ik}\rho_{i\ell}\rho_{k\ell}\right)$ で与えられる。このとき，残差 $R_{i\bullet k\ell}$ と $R_{j\bullet k\ell}$ の相関係数を求めると，次式を得る。こうして，(1−1)の拡張として偏相関係数 $\rho_{ij\bullet k\ell}$ が定義される。

(1−4)　$\rho_{ij\bullet k\ell} = \dfrac{\rho_{ij}(1-\rho_{k\ell}^2)+(\rho_{ik}\rho_{jk}+\rho_{i\ell}\rho_{j\ell})+\rho_{k\ell}(\rho_{ik}\rho_{j\ell}+\rho_{i\ell}\rho_{jk})}{\sqrt{1-(\rho_{ik}^2+\rho_{i\ell}^2+\rho_{k\ell}^2)+2\rho_{ik}\rho_{i\ell}\rho_{k\ell}}\sqrt{1-(\rho_{jk}^2+\rho_{j\ell}^2+\rho_{k\ell}^2)+2\rho_{jk}\rho_{j\ell}\rho_{k\ell}}}$

相関係数は常に $\left|\rho_{ij\bullet k\ell}\right| \le 1$ ゆえ，次の結果を導く。

<u>命題2.</u>　$\left|\rho_{ij}(1-\rho_{k\ell}^2)+(\rho_{ik}\rho_{jk}+\rho_{i\ell}\rho_{j\ell})+\rho_{k\ell}(\rho_{ik}\rho_{j\ell}+\rho_{i\ell}\rho_{jk})\right|$

$\le \sqrt{1-(\rho_{ik}^2+\rho_{i\ell}^2+\rho_{k\ell}^2)+2\rho_{ik}\rho_{i\ell}\rho_{k\ell}}\sqrt{1-(\rho_{jk}^2+\rho_{j\ell}^2+\rho_{k\ell}^2)+2\rho_{jk}\rho_{j\ell}\rho_{k\ell}}$

最後に，密度関数 $f(x_1,x_2,x_3,x_4)$ の2次形式の項は，$w_i = \sqrt{\dfrac{1-(\rho_{jk}^2+\rho_{j\ell}^2+\rho_{j\ell}^2)+2\rho_{jk}\rho_{j\ell}\rho_{k\ell}}{\det R}}z_i$ という変数変換を行うと，3次元の場合の類似式 $\sum_{i=1}^{4}w_i^2 - 2\sum_{1\le i<j\le 4}\rho_{ij\bullet k\ell}w_iw_j$ に移行することを述べて，この節を終える。

§ 2. 年齢階級別の癌罹患率の回帰分析

この節では，日本における癌罹患率の推定値 Y（[2]による）を対数変換したデータ $V=\log Y$ を分析する。説明変数としては0歳から90歳までの範囲を5年毎に区分して得られる階級値 X，およびその対数値 $U=\log X$ を用いる。(X, U, V) 3変量間の相関係数がいずれも1に近い値をとることが，本学泌尿器科学講座に属する高岡先生から教示され（その研究は2012年秋に[1]に結実する），誠に不思議に感じて V の X と U による回帰分析を勉強し始めた成果が，この節以降の内容である。次節では，実データ Y のシミュレーションモデル（[1]参照）の粗い近似をなすワイブル分布に目を向け，それに基づく V の模擬データ W を生成する。そして前節で展開した偏相関係数の議論を踏まえて，W の X と U による回帰分析を実行し，対応する V のそれと比較検討する。

さて，X と U はいずれも確定値であるが，U の X による線形回帰の残差分析を最初に行うことにより，V（次節では W）の回帰分析に適切な年齢範囲 I を確定することから始める。対応のあるデータは今，$x_k=2.5+5(k-1)\ (1\le k\le 18)$ と $u_k=\log x_k$ である。$\bar{x}=45,\ s_x=26.6927,\ \bar{u}=1.5283,\ s_u=0.4102$ そして相関係数 $r_{xu}=0.9063$（統計量を示す記法は[4]に従う）を得る。線形回帰の残差

(2−1)　$R_{u\bullet x}(k)=u_k-\bar{u}-\beta_{u\bullet x}(x_k-\bar{x}),\quad \beta_{u\bullet x}=r_{xu}\,s_u/s_x$

を求めると，次の関係式が成り立つ。

(2−2)　$$\sum_{k=1}^{18}R_{u\bullet x}(k)=0,\ \sum_{k=1}^{18}R_{u\bullet x}(k)(x_k-\bar{x})=0$$

従って平方和 $\sum_{k=1}^{18}R_{u\bullet x}^2(k)$ の自由度は16で，その分散は $s_R^2=\frac{1}{16}\sum_{k=1}^{18}R_{u\bullet x}^2(k)=0.1787^2$ と計算される。このとき，絶対値 $\left|R_{u\bullet x}(k)\right|$ の中で最大値を与えるデータ $R_{u\bullet x}(1)=-0.5384$ を s_R で割り，outlierの棄却検定（[5] p.31参照）に持ち込めば，有意水準1％でoutlierと判定される。

[2]の実データ $v_k\ (1\le k\le 18)$ については，$\bar{v}=2.2129,\ s_v=0.9340$ で，V の X による線形回帰 $(r_{xv}=0.9829)$ を構成すれば，残差 $R_{v\bullet x}(k)$ はoutlierを全く含まない。他方，V の U による線形回帰 $(r_{uv}=0.8950)$ では，残差 $R_{v\bullet u}(1)$ が10％の有意水準でoutlierと判定される。

以上，年齢の全範囲における残差分析によって，われわれは $x_1=2.5$ を捨て，5歳から90歳までの $x_k\ (2\le k\le 18)$ に範囲を限定して調べる。その結果，$\bar{x}=47.5,\ s_x=25.2488,\ \bar{u}=1.5947,$ $s_u=0.3069,\ \bar{v}=2.2855,\ s_v=0.9088,\ r_{xu}=0.9502,\ r_{xv}=0.9833,\ r_{uv}=0.9749$ を得て，線形回帰の残差 $R_{u\bullet x}(k), R_{v\bullet x}(k), R_{v\bullet u}(k)$ をそれぞれ構成すると，$R_{u\bullet x}(2)$ が5％の有意水準で，$R_{v\bullet u}(2)$ が2.5％の有意水準で，x_1 を含めた場合と同様にoutlierと判定される。

さらに，outlierを与える $x_2=7.5$ を捨て，$3\le k\le 18$ に対応する10歳から90歳までの範囲に限ると，$\bar{x}=50,\ s_x=23.8048,\ \bar{u}=1.6397,\ s_u=0.2526,\ \bar{v}=2.3807,\ s_v=0.8467,\ r_{xu}=0.9673,\ r_{xv}=0.9801,$ $r_{uv}=0.9935$ を得る。残差分析を行うと，$R_{u\bullet x}(3)$ が有意水準10％で，$R_{v\bullet u}(3)$ が有意水準2.5％で，上記と同様にoutlierと判定される。

最後に，15歳から90歳までの範囲Ⅰ（サンプルサイズ $n=15$ ）において，
3つのデータ $x_k, u_k, v_k\ (4 \le k \le 18)$ を調べると，$\bar{x}=52.5,\ s_x=22.3607,\ \bar{u}=1.6759,\ s_u=0.2143,$ $\bar{v}=2.4839,\ s_v=0.7652,\ r_{xu}=0.9769,\ r_{xv}=0.9785,\ r_{uv}=0.9980$ を得る。残差 $R_{u\bullet x}(k), R_{v\bullet x}(k), R_{v\bullet u}(k)$ はいずれも有意水準10％でのoutlierを含まないことがわかる。

偏相関係数 $r_{uv\bullet x}, r_{xv\bullet u}$ を求めると，次の表を得る。

年齢範囲	0～90	5～90	10～90	15～90
$r_{uv\bullet x}$	0.0543	0.7140	0.9023	0.9547
$r_{xv\bullet u}$	0.9110	0.8194	0.6607	0.2614

偏相関係数が1に近い（例えば15～90歳の範囲Ⅰにおける $r_{uv\bullet x}=0.9547$ ）とき，残差 $R_{v\bullet x}$ の $R_{u\bullet x}$ による線形回帰の式 $\hat{R}_{v\bullet x}(k)=\dfrac{s_{R_{v\bullet x}}}{s_{R_{u\bullet x}}} r_{uv\bullet x} R_{u\bullet x}(k)$ を作り，v の x による線形回帰の式に代入すると，

$v=\bar{v}+\dfrac{s_v}{s_x}(x_k-\bar{x})+\hat{R}_{v\bullet x}(k)$ を得る。この式は2変量 (x, u) による線形回帰の式（1−3）に一致する。
そして，残差 $R_{v\bullet xu}=R_{v\bullet x}-\hat{R}_{v\bullet x}$ が3つの関係式

$$\sum_k R_{v\bullet xu}(k)=0,\quad \sum_k R_{v\bullet xu}(k)(x_k-\bar{x})=0,\quad \sum_k R_{v\bullet xu}(k)(u_k-\bar{u})=0$$

を満たすので，その標準偏差 $s_{R_{v\bullet xu}}$ は $\sqrt{\dfrac{1}{n-3}\sum_{k=1}^{n} R^2_{v\bullet xu}(k)}$ によって求められる。

元々の s_v が，v の x による回帰直線を当てはめて，$\sqrt{(1-r_{xv}^2)\dfrac{n-1}{n-2}}$ 倍され，得られた残差 $R_{v\bullet x}$ のうち，$R_{u\bullet x}$ による回帰直線の当てはめ部分を除けば，さらに $\sqrt{(1-r_{uv\bullet x}^2)\dfrac{n-2}{n-3}}$ 倍されて $s_{R_{v\bullet xu}}$ に至る。一方，最初に v の u による回帰直線を当てはめるルートをたどれば

$\sqrt{(1-r_{uv}^2)\dfrac{n-1}{n-2}}\times\sqrt{(1-r_{xv\bullet u}^2)\dfrac{n-2}{n-3}}$ 倍という風に書ける。上記の表における $r_{uv\bullet x}$ と $r_{xv\bullet u}$ の値の著しいくい違いは，0～90歳と15～90歳とで生じている。前者では x による回帰直線の当てはめによって，v の増加の様子が大部分説明され，後者では u による回帰直線の当てはめの方が断然優位を占めることが，われわれの回帰分析の結論である。

（註2）［1］においては，25歳から75歳までの年齢範囲 Ⅰ′ をとると，V の U による回帰直線（Y で表せば，べき乗則 AX^{α} による当てはめ）がまさり，0歳から90歳までの全年齢範囲では，V の X による回帰直線（同じく Y で表せば，指数関数 $Be^{\beta X}$ による当てはめ）の方がまさることが示されている。等差数列をなす X から対数変換された U の値は，0に近づくほど変動が大きくなり，outlierが発生しやすくなるというこの節の分析結果から，［1］の主張を理解することができる。なお，V の (X, U)

による2重線形回帰の式(1−3)を Y で表せば，$\hat{Y} = CX^{\alpha}e^{\beta X}$ という風に，べき乗則と指数関数の混合形になる。

年齢範囲 I′ における x_k, u_k, x_k の統計量を記して，この節を終える。

$\bar{x} = 50,\ s_x = 15.1383,\ \bar{u} = 1.6796,\ s_u = 0.1400,\ \bar{v} = 2.4920,\ s_v = 0.5342,\ r_{xu} = 0.9909,\ r_{xv} = 0.9964,$ $r_{uv} = 0.9980$ である。残差分析を行うと I の場合と同じく，outlierに出会うことはない。

§ 3. ワイブル分布から生成される模擬データの回帰分析

t 回の細胞分裂を経たとき，ヒトが癌を発症しない確率を $S(t)$ とする。[1]で提案された発癌のシミュレーションモデルによると，$S(t)$ はポアソン近似を適用した次式で表される。

(3−1)　$S(t) = \exp\left[-\left\{\mathrm{M}\left(1-(1-p)^t\right)\right\}^{2r}\right]$

[1]のモデルを特徴づけるのは第1に，遺伝子グループの個数 r で，7以下の範囲で考察されている。そして p は変異率，M はグループ内の遺伝子数と細胞数の積から定まる正のパラメータである。

仮定　年齢 x までに起こる細胞分裂の回数 t は x の1次式である。

ところで変異率 p が極めて小さい場合，$(1-p)^t$ の2項展開のうち，p^2 以降の項を無視して，$1-(1-p)^t \fallingdotseq pt$ と近似できる。$\mathrm{M}p = \lambda, 2r = \gamma$ とおくと，(3−1)を次のワイブル分布の式([3] p.135〜9)に書き直すことができる。

(3−2)　$S(t) = e^{-(\lambda t)^{\gamma}}$

この式において，$t' = \lambda t$ は x の1次式ゆえ，S の値 $s_1 \fallingdotseq 0.99978$（この節で取り組む年齢範囲 I の始点である15歳における未発癌の推定確率）に対応する t' の値 t'_1 をとり，次に $s_{16} \fallingdotseq 0.882$（I の終点である90歳における未発癌の推定確率）に対応する t' の値 t'_{16} を求める。そして，

t' の区間 $t'_1 \le t' \le t'_{16}$ を15等分し，k 番目の分点 t'_k における S の値 $s_k = e^{-(t'_k)^{\gamma}}$ を計算して，階級値 x_k に対応する10万人当たりの癌罹患数 $(s_k - s_{k+1}) \times 10^5$ を求め，その対数を $w_k(\gamma)(k = 1,2,\cdots,15)$ とする。2から14までの偶数 γ に対し，実データ v_k と(3−2)式による模擬データ $w_k(\gamma)$ を比較研究した結果が，この節の内容である。

(註3)　[1]におけるモデルでは，2つのパラメータ p, M を別々に動かし，実データ v_k によく適合するような組 (p, M) と t を表す x の1次式を探求している。われわれのワイブル分布によるシミュレーションでは，$\mathrm{M}p = \lambda$ の形にパラメータを1元化し，横軸 t' は x の1次式と仮定できるので，λ の値を設定する必要がない。なお，(3−1)に基づいて模擬データを作成し，偏相関係数を用いた回帰分析を実行して，最適な遺伝子グループ数 r を研究するのが，著者の継続課題である。

結果1.　$w_k(\gamma)(1 \le k \le 15)$ の平均と標準偏差は次の通り。

γ	2	4	6	8	10	12	14
$\overline{w}$	2.8155	2.6427	2.5581	2.5086	2.4784	2.4575	2.4421
s_w	0.3108	0.5629	0.6504	0.6925	0.7173	0.7335	0.7450

γ とともに平均は減少し，標準偏差は増加する。

結果2.　相関係数 r_{xw}, r_{uw} および偏相関係数 $r_{uw\bullet x}, r_{xw\bullet u}$ の値は次の通り。

γ	2	4	6	8	10	12	14
r_{xw}	0.9354	0.9788	0.9897	0.9939	0.9959	0.9970	0.9977
r_{uw}	0.9886	0.9999	0.9971	0.9941	0.9918	0.9900	0.9886
$r_{uw\bullet x}$	0.9892	0.9986	0.9902	0.9823	0.9759	0.9691	0.9658
$r_{xw\bullet u}$	−0.9393	0.6419	0.9644	0.9816	0.9880	0.9909	0.9932

r_{xw} と $r_{xw\bullet u}$ は γ とともに増加する。r_{uw} と $r_{uw\bullet x}$ は $\gamma = 4$ のとき最大で，$\gamma \ge 4$ では減少する。また，$\gamma = 4$ のとき $r_{xw\bullet u}$ の絶対値が他と比べてかなり小さい点が目を引く。

結果3.　相関係数 r_{vw} および偏相関係数 $r_{vw\bullet x}, r_{vw\bullet u}, r_{vw\bullet xu}$ の値は次の通り。

γ	2	4	6	8	10	12	14
r_{vw}	0.9833	0.9985	0.9971	0.9947	0.9927	0.9911	0.9898
$r_{vw\bullet x}$	0.9315	0.9633	0.9715	0.9745	0.9759	0.9774	0.9768
$r_{vw\bullet u}$	−0.3435	0.6296	0.4128	0.3781	0.3598	0.3513	0.3392
$r_{vw\bullet xu}$	−0.2964	0.6237	0.6295	0.6715	0.6806	0.7110	0.7096

r_{vw} と $r_{vw\bullet u}$ は $\gamma = 4$ のとき最大で，$\gamma \ge 4$ では減少する。他方，$r_{vw\bullet x}$ と $r_{vw\bullet xu}$ は $\gamma = 12$ のとき最大で，$\gamma \le 12$ では増加する。なお，$r_{vw}, r_{vw\bullet x}$ の値がどれも1に近い点が目を引く。

謝辞

§1で展開した偏相関係数の議論は，本学医療情報部木村先生からの鋭い質問がきっかけを与えてくれた。また，§2の高い相関関係をもつ (*X*, *U*, *V*) 3変量の研究は，本学泌尿器科学講座の高岡先生に負っている([1])。両先生に感謝の言葉を記す次第です。最後に，物理学実験の多忙な月日にも関わらず，原稿完成に向けて助けて下さった吉田(鴨藤)江利子さんに，心から御礼を申しあげます。

参考文献

[1] Takaoka N, Noda A, Takayama T, Ozono S: A multi-hit gene group model based on cell division closely simulates actual cancer incidence in Japan and the United States. British J of Cancer (submitted).

[2] Matsuda T, Marugame T, Kamo K, Katanoda K, Ajiki W, Sobue T: Cancer incidence and incidence rates in Japan in 2006: based on data from 15 population-based cancer registries in the monitoring of cancer incidence in Japan (MCIJ) project. Jpn J Clin Oncol **42**: 139-47, 2012.

[3] Lee ET: Statistical methods for survival data analysis. New York: John Wiley & Sons, Inc., 1992.

[4] 東京大学教養学部統計学教室（編）：統計学入門．東京大学出版会，1991．

[5] 統計数値表編集委員会（編）：簡約統計数値表．日本規格協会，1977．