多変量解析 ver3.0

KENZOU

2009 年 7 月 1 日

前書き

統計的手法でデータを処理するといったシーンはいろんなところで出てきますね。もともとこの小稿をまとめるきっかけになったのは，小生の関係するある若き医療従事者が“ 統計的な手法でデータをまとめ，有意な結論をだしていかないと。。。統計は独学してもよく分からないし，なにか適当な講義を聴きに行こうかなと思っているの ”というのを聞き，それなら少しでも役に立てればということで，内緒でボソボソとまとめ始めたのがきっかけ。後日，その若き医療従事者に“ 多変量解析を纏めつつあるんだけど ”と言うと，“ 何それ？”との返事。“ いや〜統計でデータ解析云々といってたので，丁度いいかなと思って多変量解析の話を纏め始めたんだけど… ” “ ふ〜ん… まっボチボチやるから今は特にいいの ”というご返答。ヤレヤレ先走りしてやるもんではないワイと内省しきり（爆；）。しかし，いずれにしてもやり始めたからにはある程度格好のつくところまで進めなくては気がすまないという難儀な性格なので，ともかく継続モードに入る。

単回帰・重回帰分析までが ver1.0 (09/03/14)，主成分分析を加えたのが ver2.0 (09/04/14)，そしてしばらく間をおいて因子分析の稿を付加して今回の ver3.0 と相成る。因子分析の数学的な取り扱いは結構面倒な面があり，データを与えられて手計算で因子分析をしようとはじめると一発で撃沈してしまう（爆；）。しかし，適当な統計解析ソフトを使えば，何の苦痛もなしに即座に答えを出してくれる。それなら，最初からそうすればいいじゃん，何もややこしい数学理論まで踏み込まなくても。。。となりそうだが，どっこい答えの解釈にはどうしても因子分析の知識が必要になる。このため，ある程度はその中身に踏み込んでおかないと。

この分野に興味のある方のために，参考にしたテキストを紹介すると，涌井良幸，涌井貞実・共著「図解で分かる多変量解析」（日本実業出版社），有馬哲，石村貞夫・共著「多変量解析のはなし」（東京図書），柳井久江著「エクセル統計-実用多変量解析編」（OMS 出版）（←付録としてエクセル・アドインソフトの「Mulcel」がついている。因子分析の実際の計算はこのソフトを活用した）等といったところ。ネット上には残念ながら参考になる資料が殆ど見当たらなかった。

さて，あと「正準相関分析」と「判別分析」を予定しているが，脱稿はいつになることか本人もよく分からない (爆；)。熱い request をいただいてケツをひっぱ叩かれるか，本人がその気になるまで取り掛からない。

拙稿でわけのわからないところにぶつかれば，適当な参考書を紐解いて追求していただきたい。また，浅学の身ゆえ，誤りや誤解の記述が多くあろうかと思う。それらを見つけられれば，お手数でも一報いただけると嬉しい。

目 次

1　多変量解析の準備

1.1　平均・分散・標準偏差

1.1.1　平均と分散・標準偏差

n 個のデータ $x_1, x_2, \cdots, x_n$ からなる標本（サンプル）[1]の**平均** $\bar{x}$ は

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{1.1}$$

で与えられます。仮にサンプルの平均が同じであっても，下の表の【A】,【B】,【C】にみるように平均の周りのデータのバラツキまで同じとは限りません。平均からのバラツキの程度を測る尺度として（1.2）式で表される**分散**というものを定義します。各データの平均からのズレを足し合わたものをデータの数 n で割れば平均的なバラツキが把握できると考えられますが，バラツキは平均を中心にプラスとマイナスが均等に分布しているため，その結果は 0 となる[2]のでうまくありません。そこで，ズレを 2 乗してすべてプラスの値とし，それらを足し合わせたものを $n-1$ で割ることにするわけです。分散を s_x^2 と書くと

【A】

No	x
1	50
2	50
3	50
4	50
5	50
6	50
7	50
8	50
9	50
10	50
平均	50
分散	0
標準偏差	0
変動係数	0

【B】

No	x
1	60
2	45
3	50
4	30
5	60
6	50
7	40
8	45
9	70
10	50
平均	50
分散	128
標準偏差	11.3
変動係数	0.226

【C】

No	x
1	20
2	45
3	50
4	95
5	80
6	55
7	5
8	15
9	50
10	85
平均	50
分散	928
標準偏差	30.5
変動係数	0.609

$$\text{分散}：s_x^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{1.2}$$

で定義されます。平均からのズレを**偏差**といい，その 2 乗の総和 $\sum_{i=1}^{n}(x_i-\bar{x})^2$ を**偏差平方和**といいます。分散はデータのバラツキを見る尺度としてはいいのですが，2 乗の計算をしているので単位が異なってしまうという問題が生じます。元のデータが身長のように cm という単位であれば，分散の単位は cm^2 と面積の単位になってしまいます。そこで元の単位とあわすためには分散の平方根をとればいいわけで，それを**標準偏差** s_x と呼び，次式で定義されます。

$$\text{標準偏差}：s_x = \sqrt{s_x^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{1.3}$$

また，標準偏差を平均で割った**変動係数**というものがあります。例えば互いに異なる平均をもつ 2 つの集団があり，そのデータのバラツキ具合を比較したいとします。バラツキ具合として標準偏差を比較すれば良いように思われますが，平均が異なるとそれに比例して標準偏差も異なるので単純に比較できません。そこで，平均が異なっていてもバラツキ具合が比較できるように標準偏差を平均で割り，平均と無関係になった値（単位のない無名数）で比較すればよいことになります。これを変動係数と呼んでいます。

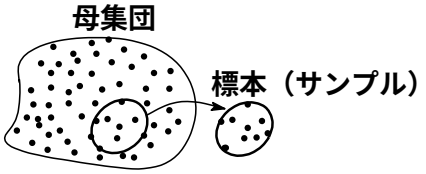
$$\text{変動係数} = \frac{s_x}{\bar{x}} = \frac{\text{標準偏差}}{\text{平均}} \tag{1.4}$$

---

[1] 集団全体を母集団といい，これから一部を抜き取った要素を標本（サンプル）といいます。

[2] 簡単のために $n=3$ の場合を調べると $\{x_1-(x_1+x_2+x_3)/3\}+\{x_2-(x_1+x_2+x_3)/3\}+\{x_3-(x_1+x_2+x_3)/3\}=0$。なぜ n で割らないのか，それは母集団から採取した標本データの場合は $n-1$ で割ることになっているからです。母集団を対象としている場合は n で割ります。詳しいことは適当な統計学のテキストを参照ください。

この値が大きくなるほど変動が大きいといえます。一般に 0.2 を区切りとして，0.2 を超えると変動大，それ以下であれば変動小と判断されます。

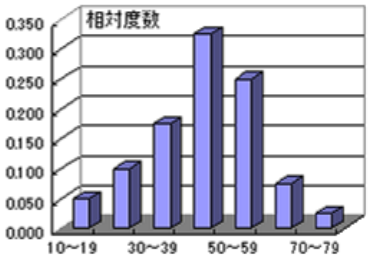


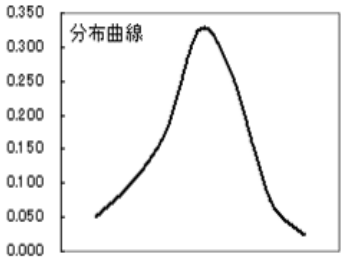
1.2 正規分布

1.2.1 度数分布

あるクラスの英語のテストの成績を得点を 10 点ごとに区切ってその中に入る人数を整理すると下表左のようになったとします。この区切りを**階級**といい，それぞれの階級に入るデータの数を**度数**，この表を**度数分布表**と呼んでいます。この例では 7 階級の度数分布ということになります。また，各階級の度数が全度数に占める割合を**相対度数**といいます。度数分布表をグラフにしたものを**ヒストグラム**（中央の図）といいますが，階級の幅をどんどん小さくしていくとヒストグラムの輪郭は滑らかな曲線になっていきます。これを**分布曲線**と呼んでいます。

点数(階級)	人数(度数)	相対度数
10〜19	2	0.050
20〜29	4	0.100
30〜39	7	0.175
40〜49	13	0.325
50〜59	10	0.250
60〜69	3	0.075
70〜79	1	0.025





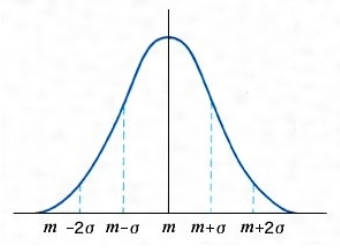
1.2.2 正規分布

分布曲線の中で最も代表的なものは，左右対称な釣鐘型の正規分布と呼ばれるもので

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} \tag{1.5}$$

で表されます。慣例に従って標準偏差を σ で表しました。正規分布の分散は σ^2 ですね。

右の曲線で中央の m が平均 $\bar{x}$ を表し，標準偏差 σ は曲線の中腹辺り（変曲点）を表しています。(1.5) より正規分布は平均と分散（または標準偏差）が決まると完全に形が決まります。正規分布に従う x について次のことが言えます。



- $m - \sigma \leqq x \leqq m + \sigma$ となる x 全体の 68.2%ある
- $m - 2\sigma \leqq x \leqq m + 2\sigma$ となる x 全体の 95.4%ある
- $m - 3\sigma \leqq x \leqq m + 3\sigma$ となる x 全体の 99.2%ある

1.3 相関係数と共分散

ある学校の生徒 10 人の身長 x と体重 y を測定したところ，下表のようになったとします。身長や体重のようにデータの“ ラベル ”となるようなものを**変量**といいます。身長と体重のデータを xy

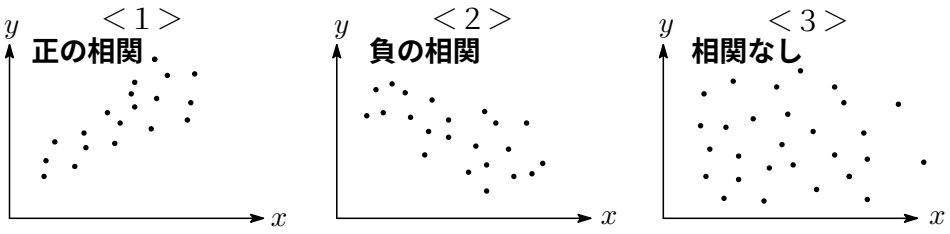
平面上にプロットしたものを**散布図**といいます。
散布図を見ると x と y の関係，つまり身長と体重の間に関係があるのかないのか，いわゆる**相関関係**を見ることができます。散布図からわかる相関関係としてつぎの 3 つのパターンがあります。

(1) 正の相関： x が増加（減少）すると y も増加（減少）する

(2) 負の相関： x が増加（減少）すると y は減少（増加）する

(3) 相関なし： x と y の間になんの傾向も認められない

No	身長(x)	(A) x−xi	体重(y)	(B) y−yi	A×B
1	146	−4	45	−5	20
2	145	−5	46	−4	20
3	147	−3	47	−3	9
4	149	−1	49	−1	1
5	151	1	48	−2	−2
6	149	−1	51	1	−1
7	151	1	52	2	2
8	154	4	53	3	12
9	153	3	54	4	12
10	155	5	55	5	25
共分散					10.89
平均	150		50		
標準偏差	3.40		3.50		



2 つの変量の間に因果関係があれば必ず相関関係は存在しますが，相関関係があるからといって必ずしも因果関係が存在するとは言えません。2 つの変量の動きが別の原因に依存しているのも拘わらず相関があるかのように見える“ 偽りの相関 ”もあるので注意が必要です。

1.3.1 相関係数

散布図からは正の相関，負の相関，ゼロの相関という相関関係の“ 方向性 ”（上向きか下向きか向き無しか）を読み取ることができました。次ぎに相関の強弱を示す尺度として**相関係数**（correlation cofficient) というものを定義します。また，相関係数の符号から相関関係の方向性も分かります。いま，n 個のデータの組 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ があった場合，変量 x と y の相関係数を r_{xy} とすると

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} = \frac{\text{積和}}{\sqrt{x \text{ の偏差平方和} \times y \text{ の偏差平方和}}} \tag{1.6}$$

で定義されます。r_{xy} のとりうる範囲を調べるために $\xi_i = x_i - \bar{x}$, $\eta_i = y_i - \bar{y}$ とおくと（1.6）は，

$$r_{xy}^2 = \frac{(\sum \xi_i \eta_i)^2}{\sum \xi_i^2 \sum \eta_i^2}$$

ところで[3]

$$\sum \xi_i^2 \sum \eta_i^2 - \left(\sum \xi_i \eta_i\right)^2 = \sum (\xi_i \eta_j - \xi_j \eta_i)^2 \geqq 0 \longrightarrow \sum \xi_i^2 \sum \eta_i^2 \geqq \left(\sum \xi_i \eta_i\right)^2$$

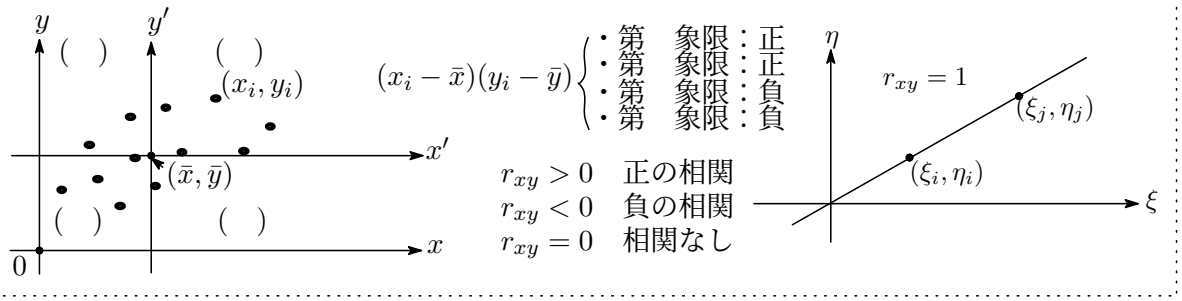
となるので $r_{xy}^2 \leqq 1 \longrightarrow -1 \leqq r_{xy} \leqq 1$ となります。

■相関の方向性

相関係数 r_{xy} から相関の方向性がでてくるのを以下に見ていきます。(1.6) の分子 $\sum (x_i - \bar{x})(y_i - \bar{y})$ は変量 x, y の平均からの偏差の積なので，平均の点 $(\bar{x}, \bar{y})$ を原点とする新しい座標系で $(x_i - \bar{x})(y_i - \bar{y})$ の符号を調べると，第　，第　象限では正，第　，第　象限では負となります。したがって多く

[3] 簡単のため $i = 2$ の場合を見ると $(\xi_1^2 + \xi_2^2)(\eta_1^2 + \eta_2^2) - (\xi_1\eta_1 + \xi_2\eta_2)^2 = (\xi_1\eta_2 - \xi_2\eta_1)^2$

の点が第　，　象限に分布しているような正の相関では偏差の積の和 $\sum(x_i-\bar{x})(y_i-\bar{y})$ は正の値をとり，多くの点が第　，　象限に分布している負の相関では負の値をとることになるので，r_{xy} の符号が相関の方向性を示します。



尚，データがが各象限に一様に分布している場合は，各偏差の積 $(x_i-\bar{x})(y_i-\bar{y})$ の値は正値も負値も一様に出現するので，それらを足し合わせると $\sum(x_i-\bar{x})(y_i-\bar{y})=0$ となり，この場合は相関無しということになります。

■相関の強さ

次に相関の強さですが，$(x_i-\bar{x})(y_i-\bar{y})$ は新しい座標系の各点で描かれる長方形の面積を意味し，r_{xy} の分子 $\sum(x_i-\bar{x})(y_i-\bar{y})$ はこれら面積の総和を表します。したがって，正の相関関係が存在する場合，$\sum(x_i-\bar{x})(y_i-\bar{y})$ の値が大きな値をとればとるほど原点 $(\bar{x},\bar{y})$ から遠くに分布する点がより多く存在することを意味し，強い右肩上がりの傾向を示すことになります。一方，負の相関の場合は，$(x_i-\bar{x})(y_i-\bar{y})$ が負の値となりますが，その総和の絶対値 $|\sum(x_i-\bar{x})(y_i-\bar{y})|$ が大きければ大きいほど強い右肩下がりの傾向を示すことになります。ただし，$\sum(x_i-\bar{x})(y_i-\bar{y})$ は単にデータの数が増えるだけでもそれに比例して大きくなるのでそれをなんとかしなくてはいけない。そこでデータの個数（実際は n ではなく $n-1$）で割ってやることにする[4]。これを x と y の**共分散**と呼び，次式で定義します。

$$s_{xy}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \tag{1.7}$$

これで解決したかと思われますが，これもデータの桁が大きくなると，当たり前ですが $(x_i-\bar{x})(y_i-\bar{y})$ も大きな値となってしまう。そこで変動係数のところでやったように同じ桁数の標準偏差で割ってやればよいということで，結局，相関係数は

$$\begin{aligned} r_{xy}&=\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}} \\ &=\frac{s_{xy}}{s_x\cdot s_y} \end{aligned} \tag{1.8}$$

と表せばよいことになります。$r_{xy}^2=1$ の場合は、等式 (1.7) よりすべての $i,\,j$ に対して $\xi_i\eta_j-\xi_j\eta_i=0\longrightarrow\frac{\xi_i}{\eta_i}=\frac{\xi_j}{\eta_j}$ が成立するときに限ります。これはすべての点 $(\xi_1,\eta_1),(\xi_2,\eta_2),\cdots,(\xi_n,\eta_n)$ が原点を通る同一直線上にのっていることになり x_i と y_i は完全な相関をもちます。また、$-1<r_{xy}<1$ の場合では、$r_{xy}>0$ ならば正の相関，$r_{xy}<0$ ならば負の相関をもちます。$r_{xy}=0$ の場合は、変量 $x,\,y$ 間に全く相関がないことになります。r_{xy} の値と相関の強さは，次のようになります。

$$\left.\begin{array}{ll} 0.9\leqq r_{xy}<1 & \text{非常に強い相関} \\ 0.7\leqq r_{xy}<0.9 & \text{やや強い相関} \\ 0.5\leqq r_{xy}<0.7 & \text{やや弱い相関} \end{array}\right\}\ \text{相関あり}$$

$$r_{xy}<0.5 \qquad \text{非常に弱い相関}\quad\cdots\ \text{相関なし}$$

[4] n でなく $n-1$ とする理由は脚注 2 を参照。

＜メモ：相関係数をベクトル的に考察すると＞

次のベクトル $\boldsymbol{a}$，$\boldsymbol{b}$を考えます。

$$\boldsymbol{a} = (x_1 - \bar{x}, x_2 - \bar{x}, \cdots, x_n - \bar{x}), \quad \boldsymbol{b} = (y_1 - \bar{y}, y_2 - \bar{y}, \cdots, y_n - \bar{y})$$

ベクトル $\boldsymbol{a}$, $\boldsymbol{b}$ のなす角を θ とすると、ベクトルの内積は

$$\begin{aligned}
&\boldsymbol{a} \cdot \boldsymbol{b} = |\,\boldsymbol{a}\,| \cdot |\,\boldsymbol{b}\,| \cos\theta \\
\therefore \quad &\cos\theta = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\,\boldsymbol{a}\,| \cdot |\,\boldsymbol{b}\,|} \\
&= \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2} \cdot \sqrt{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}} \\
&= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \cdot \sqrt{\sum (y_i - \bar{y})^2}}
\end{aligned}$$

この式の右辺は相関係数（1.8）の右辺と同じですね。つまり、相関係数 r_{xy} は 2 つのベクトル $\boldsymbol{a}$, $\boldsymbol{b}$ のなす角の余弦（$\cos\theta$）を表していることになります。相関係数 r_{xy} が 1 とは，$\cos\theta = 1$ つまり $\theta = 0\,^\circ$で 2 つのベクトルは同じ方向を向いており、相関係数が 0 とは $\cos\theta = 0\,^\circ$つまり $\theta = 90\,^\circ$で 2 つのベクトルは直交しているということになります。

さて，最初にあげた生徒の身長と体重の相関関係を調べて見ましょう。

・共分散 $s_{xy} = \dfrac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y}) = 10.89$

・標準偏差 $s_x = 3.40$, $s_y = 3.50$

・相関係数 $r_{xy} = \dfrac{s_{xy}}{s_x s_y} = 0.915$

相関係数 0.915 が得られたので，身長と体重には大変強い相関があることになります。

【例題】次の表は 6 つの営業所の売上額，広告費，営業担当者数，キャンペーンの実施有無を示したものです。このデータにおける変数相互の相関係数を求めてみます。なお，キャンペーンは「有り」を 1,「無し」を 0 として数量データに置き換えることにします[5]。相関係数は（1.6）の計算から得られますが，ここではエクセルの分析ツール[6]で相関係数を求めることにします。

	売上額	広告費	営業担当者数	キャンペーン
A	8	5	6	0
B	9	5	8	1
C	13	7	10	1
D	11	5	11	0
E	14	8	12	0
F	17	12	13	1

「ツール」→「分析ツール」で「相関」を選択します。入力元として「入力範囲」を求めてきますのでマウスで“ 売上額 ”から“ キャンペーン－Ｆ ”までの領域をクリック選択し,「先頭行をラベルとして使用 (L)」にチェックを入れます。出力オプションで「新規ブック」のラジオボタンを選択して「ＯＫ」をクリックします。その結果は

[5] キャンペーンの有無のような数量データでないデータをカテゴリデータといいます。

[6] 分析ツールはエクセルのアドインソフトとなっており，登録はメニューの「ツール（T) →アドイン（T) →「分析ツール」をチェック→ OK をクリックで登録できます。

	売上額	広告費	営業担当者数	キャンペーン
売上額	1			
広告費	0.932143217	1		
営業担当者数	0.916698497	0.751160094	1	
キャンペーン	0.327326835	0.397359707	0.140028008	1

が得られます。売上額と広告費，売上額と営業担当者数の相関係数はそれぞれ 0.932, 0.917 と大変強い相関があることが分かります。一方，売上額とキャンペーンはあまり強い相関がなく，売上げを増すには広告費や営業マンを増やすべきであるという結論が導けます。

1.4 標準化

標準偏差はバラツキを表す指標でした。セクション 1.1.1 の【B】と【C】のデータは平均は同じでも標準偏差が異なっていました。つまり，集団としての平均は同じでもバラツキ具合は異なるということです。そこで，個々のデータのバラツキ比較をしたい場合はどうすればよいかということですが，それぞれの偏差を標準偏差で割ったもの（個々のバラツキを全体のバラツキで割ったもの）を比較すればよいということになります。これをデータの**標準化**とか**基準化**といっています。

【B】

No	×	$x-\bar{x}$	$(x-\bar{x})/\sigma$
1	60	10	0.88
2	45	−5	−0.44
3	50	0	0.00
4	30	−20	−1.77
5	60	10	0.88
6	50	0	0.00
7	40	−10	−0.88
8	45	−5	−0.44
9	70	20	1.77
10	50	0	0.00
平均	50	0	0.00
分散	128	128	1.00
標準偏差	11.3		
変動係数	0.226		

【C】

No	×	$x-\bar{x}$	$(x-\bar{x})/\sigma$
1	20	−30	−0.98
2	45	−5	−0.16
3	50	0	0.00
4	95	45	1.48
5	80	30	0.98
6	55	5	0.16
7	5	−45	−1.48
8	15	−35	−1.15
9	50	0	0.00
10	85	35	1.15
平均	50	0	0.00
分散	928	928	1.00
標準偏差	30.5		
変動係数	0.609		

$$\text{標準化：}\quad z_i = \frac{x_i - \bar{x}}{s_x} \tag{1.9}$$

上表の右側には標準化したデータが載っています。このように標準化すると，z の単位は無次元なのでいろいろな単位で表された資料間での単純比較が可能になる。

さて，標準化されたサンプルは，平均が 0，分散は 1 を示すことになります。

$$\begin{cases} \text{平均：}\quad \bar{z} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} z_i = \dfrac{1}{n}\sum_{i=1}^{n}\left(\dfrac{x_i - x}{s_x}\right) = \dfrac{1}{n\,s_x}\left(\sum_{i=1}^{n} x_i - n\bar{x}\right) = 0 \\ \text{分散：}\quad s_x^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(z_i - \bar{z})^2 = \dfrac{1}{n-1}\sum_{i=1}^{n} z_i^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(\dfrac{x_i - \bar{x}}{s_x}\right)^2 = 1 \end{cases}$$

また，z の符号や大小により個々のデータの特長を次のように表すことができます。

(1) z が正なら標準より大きく，負なら標準より小さい。

(2) z の大きさが 1 より大きければ，標準より大きく離れている（バラツイている）。

1.5 単回帰分析

x と y の間になんらかの相関関係が見られる場合，適当な直線を引いてやれば，x に対する y の平均的な値を推定することができます。このような直線を回帰直線と呼んでいます。

$$\text{回帰直線：} y = ax + b$$

いま，身長（x) と体重（y）の 2 変量について n 人分のデータが得られ，身長 x から体重 y を予測する式を作りたいとします。まず，a, b を未知数として予測式を

$$Y = ax + b \tag{1.10}$$

としましょう。予測値を実測値と区別するために大文字の Y で書くことにします。各データ（x_i, y_i）に対して回帰直線（1.10）から得られる予測値を Y_i とすると

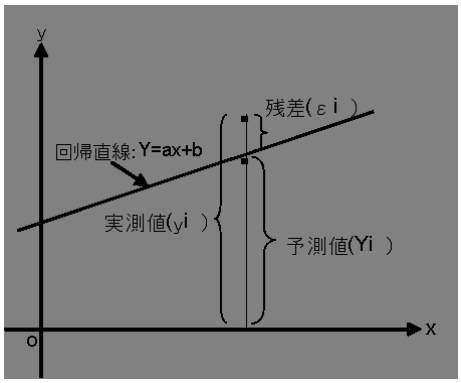
$$Y_i = ax_i + b \tag{1.11}$$

実測データと予測値とのズレ（残差）を ε と書くと

$$\varepsilon_i = y_i - Y_i = y_i - (ax_i + b) \tag{1.12}$$

と表されます。

求める回帰直線は残差 $\varepsilon_i\,(i = 1, 2, \cdots, n)$ が最小となるような a, b を決めることから求まります。ところで残差はプラスとマイナスがあり，必要なのは残差の大きさ（絶対値）を最小にすることですから，残差を 2 乗した総和，**残差平方和**を最小にするような定数 a, b を求めることにします。この方法を**最小二乗法**（least square method）といいます。



1.5.1 最小二乗法

残差平方和を S とすると

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2 = \sum_{i=1}^{n} (y_i - Y_i) \tag{1.13}$$

最小二乗法は S の極値問題となるので，S を変数 a, b でそれぞれ偏微分した式を 0 とおき，この連立方程式を解いて a, b を求めればよいことになります。

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\displaystyle\sum_{i=1}^{n}(y_i - ax_i - b)x_i = 0 \longrightarrow \left(\sum_{i=1}^{n} x_i^2\right) a + \left(\sum_{i=1}^{n} x_i\right) b = \sum_{i=1}^{n} x_i y_i \\ \dfrac{\partial S}{\partial b} = -2\displaystyle\sum_{i=1}^{n}(y_i - ax_i - b) = 0 \quad \longrightarrow \left(\sum_{i=1}^{n} x_i\right) a + nb = \sum_{i=1}^{n} y_i \end{cases} \tag{1.14}$$

連立方程式の解はクラメルの公式（Cramer's formula）より

$$\begin{cases} Ax + By = p \\ Cx + Dy = q \end{cases} \longrightarrow x = \frac{\begin{vmatrix} p & B \\ q & D \end{vmatrix}}{\begin{vmatrix} A & B \\ C & D \end{vmatrix}}, \quad y = \frac{\begin{vmatrix} A & p \\ C & q \end{vmatrix}}{\begin{vmatrix} A & B \\ C & D \end{vmatrix}} \quad (\text{但し } AD - BC \neq 0)$$

となるので，連立方程式（1.14）の解は

$$A \to \sum_{i=1}^{n} x_i^2, \quad B \to \sum_{i=1}^{n} x_i, \quad p \to \sum_{i=1}^{n} x_i y_i$$
$$C \to \sum_{i=1}^{n} x_i, \quad D \to n, \quad q \to \sum_{i=1}^{n} y_i$$

と置き換えてクラメルの公式を使うと[7]

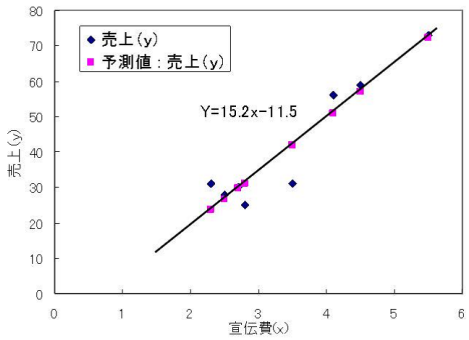
$$\begin{cases} a = \dfrac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum {x_i}^2 - (\sum x_i)^2} \\ b = \dfrac{\sum x_i^2 \sum y_i - \sum x_i y_i \sum x_i}{n\sum {x_i}^2 - (\sum x_i)^2} \end{cases} \tag{1.15}$$

と得られます。

それでは，早速次の例をやって見ましょう。ある販売会社の営業所別宣伝費（単位百万円）と売上（単位千万円）の関係データが下表のようになっていたとします。これから売上高と宣伝費の単回帰分析をやってみましょう。

営業所	宣伝費 (x)	売上（y）
1	5.5	73
2	4.5	59
3	4.1	56
4	3.5	31
5	2.5	28
6	2.3	31
7	2.7	30
8	2.8	25

売上高 (y) は宣伝費 (x) の一次式 $y = ax + b$ で表されると仮定して係数 a, b を求めます。係数 a, b は（1.15）を使えば求まりますが，ここでは手間を省いてエクセルで求めることにします。メニューの「ツール」→「分析ツール」→「回帰分析」をクリックし,「入力 Y 範囲」（売上げの列）「入力 X 範囲」（宣伝費の列）をマウスで領域選択して入力し,「ラベル」をチェックします。「出力オプション」で「新規ブック」を選択し,「観測値グラフの作成」をチェックして「OK」をクリックするとズラッといろいろな結果が表示されます。ここで必要なのは“ 切片 ”と“ 宣伝費 (x) ”の“ 係数 ”で切片：−11.5199496，宣伝費 (x)：15.23869523 が得られます。これから売上げと宣伝費をの関係を結びつける回帰直線の方程式は



$$Y = 15.2x - 11.5$$

と得られます。

1.5.2　回帰方程式と分散・共分散の関係について

回帰方程式の特長を見てみましょう。（1.14）の第 2 式の両辺をデータの数 n で割って整理すると

$$\left(\sum_{i=1}^{n} x_i\right) a + nb = \sum_{i=1}^{n} y_i \longrightarrow \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) a + b = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\therefore \quad \bar{y} = a\bar{x} + b \tag{1.16}$$

[7] $\sum_{i=1}^{n} \longrightarrow \sum$ で表しています。

となりますが，これは傾きが a で点（$\bar{x}$, $\bar{y}$）を通る直線の方程式ですね。つまり，回帰直線は変量の平均，いい換えると分布の中心（重心）を通る直線であることになります。次ぎに回帰直線の傾き a の中身を詳しく見ていきましょう。(1.16）より

$$b = \bar{y} - a\bar{x}$$

これを（1.13）に入れて S の最小条件を適用すると

$$S = \sum_{i=1}^{n} \{(y_i - \bar{y}) - a(x_i - \bar{x})\}^2 \longrightarrow \frac{\partial S}{\partial a} = 2\sum_{i=1}^{n} \left\{(x_i - \bar{x})(y_i - \bar{y}) - a(x_i - \bar{x})^2\right\} = 0$$

$$\therefore \quad \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - a\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0 \tag{1.17}$$

(1.17) の両辺を $n-1$ で割ると

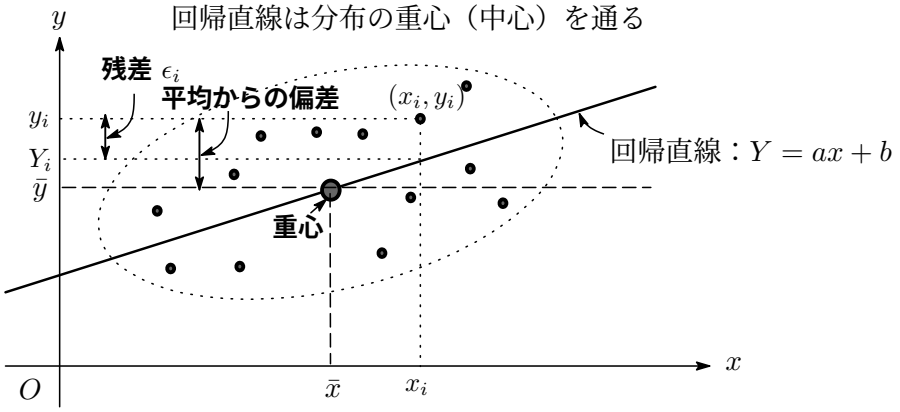
$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = a\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \longrightarrow s_{xy} = as_x^2$$

$$\therefore \quad a = \frac{s_{xy}}{s_x^2} = \frac{s_y}{s_x}r_{xy} \quad (\text{相関係数：} r_{xy} = \frac{s_{xy}}{s_x s_y}) \tag{1.18}$$

となって，回帰直線の傾き a は分散・共分散上式のように相関係数 r_{xy} に比例していることが分かります。（1.18）を回帰直線 $y = ax + b$ に入れると

$$y = \frac{s_{xy}}{s_x^2}x + (\bar{y} - a\bar{x}) = \frac{s_{xy}}{s_x^2}\,x + \left(\bar{y} - \frac{s_{xy}}{s_x^2}\,\bar{x}\right) \tag{1.19}$$

となります。また,（1.18）より s_x, s_y が 1 のときは，回帰直線の傾き a は相関係数 r_{xy} に等しくなります。



1.5.3 決定係数（寄与率）

得られた回帰方程式はどのくらい当てはまり具合が良いのか，これを判断する指標として決定係数というのがあります。以下，それを見ていきます。さて，目的変量の実測値の分散（s_y^2）と予測値の分散（s_Y^2），残差の分散（s_ε^2）の間に次の関係式が成立します。

$$s_y^2 = s_Y^2 + s_\varepsilon^2 \tag{1.20}$$

ただし，

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2, \quad s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{y})^2, \quad s_\varepsilon^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - Y_i)^2$$

実測値の分散と予測値，残差の分散を関係付ける（1.20）は次のようにして証明できます。

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}\{(y_i-Y_i)+(Y_i-\bar{y})\}^2=\sum_{i=1}^{n}(y_i-Y_i)^2+\sum_{i=1}^{n}(Y_i-\bar{y})^2+2\sum_{i=1}^{n}(y_i-Y_i)(Y_i-\bar{y})$$

$$\sum_{i=1}^{n}(y_i-Y_i)(Y_i-\bar{y})=\sum_{i=1}^{n}\{(y_i-\bar{y})-a(x_i-\bar{x})\}\{a(x_i-\bar{x})\}$$

$$=a\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})-a^2\sum_{i=1}^{n}(x_i-\bar{x})^2=a\{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})-a\sum_{i=1}^{n}(x_i-\bar{x})^2\}$$

$$=0\quad\left(\because\ a=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}\right)$$

$$\therefore\quad\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(Y_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-Y_i)^2\longrightarrow s_y^2=s_Y^2+s_\varepsilon^2$$

Q.E.D

実測値の分散 s_y^2 はデータが与えられた時点で一定の値をとるので，関係式より予測値と残差の各分散の和は一定ということになります。最小二乗法は残差の分散を最小にする手法でしたが，このことは言い換えると 予測値の分散を最大にする ということにつながります。そこで，予測値の分散を実測値の分散で割ったものを考えると，これは残差の分散が 0 のとき最大値 1 をとるので，回帰方程式の当てはまりの良さを表す尺度として最適です。この値を**決定係数**または**寄与率**と呼び，R^2 で表します。

$$R^2=\frac{\text{目的変量の予測値の分散}}{\text{目的変量の実測値の分散}}=\frac{s_Y^2}{s_y^2}\tag{1.21}$$

ちなみに、決定係数の平方根 R を**重相関係数**といいますが，これは後ほど重回帰分析のところででてきます。（1.20）より

$$R^2=1-\frac{s_\varepsilon^2}{s_y^2}\tag{1.22}$$

となり，定義より R^2 のとりうる範囲は

$$0\leqq R^2\leqq 1\tag{1.23}$$

となります。$s_\varepsilon=0$ のときは予測値（Y_i）と実測値（y_i）が等しくなり，$R^2=1$。この場合は回帰方程式の当てはまり具合が最も良いことになります。逆に，当てはまり具合が最も悪い場合は予測値の分散が 0，つまり予測値（Y_i）が目的変量（y）の平均に等しくなるときで，(1.20）より $s_y^2=s_\varepsilon^2$ で $R^2=0$ となります。決定係数 R^2 の値と当てはまりの精度は以下のようです。

R^2 の値	精 度
0.8 以上	非常に良い
0.5 以上	良い
0.25 以上	やや良い
0.25 未満	良くない

ちなみに，先ほどのある販売会社の営業所別宣伝費と売上の資料の決定係数はエクセルの計算結果 $R^2=0.895$ が得られ，回帰方程式 $Y=15.2x-11.5$ は非常に良い精度で成り立っているということになります。

《注意 !》決定係数が大きいから良い回帰方程式と単純に捉えるのは危険で，決定係数 R^2 は説明変量の数を増やすとそれに伴って増加してしまうという性質を持っています。だから目的変量にあまり影響を及ぼすとは考えられない説明変量を付け加えても決定係数は大きくなるので，R^2 が 1 に近いからといって良い回帰方程式であるとはいい切れません。決定係数が大きいことは良い回帰方程式であるための必要条件

であって十分条件ではありません。良い回帰方程式であるか否かは，重回帰式の分散分析により“ この重回帰式は目的変量 y の予測に役立つか ”という検定を行う必要がありますが，詳しいことは適当なテキストを参照ください。

2　多変量解析

多変量解析というのは多くの変量データを解析するための手法で，各変量相互の関係をある関係式で捉え，相互関係を分析する手法のことです。具体的には，重回帰分析，主成分分析，判別分析，因子分析等々の手法があり，これらを総称して多変量解析と呼んでいます。いろいろな現象はいくつかの要因が複雑に絡み合ってできていますが，それら要因の関係を解きほぐし，現象を解明する一つの有力な武器といえます。

さて，理論的な話は後回しにして，具体的な例でその感触を掴んでいきましょう。ある販売会社の 6 営業所における宣伝費（x)，営業マンの数（u)，売上額（y）次のようになっていたとします。

	宣伝費 (x)[百万円]	**営業マンの数**(u)	**売上額** (y)[千万円]	**回帰方程式より** (y)
A	5	6	8	8.1
B	5	8	9	9.4
C	7	10	13	12.0
D	5	11	11	11.3
E	8	12	14	14.0
F	12	13	17	17.4
G	13	14	?	

上の表をみると，宣伝費や営業マンが多いほど売上額が多いことが分かります。そこで新設する G 営業所は宣伝費と営業マンの数を他の営業所より多めに設定しました。果たして G 営業所の売上額はどれだけ見込めるでしょうかという問題で，A〜F の営業所のデータに多変量解析（重回帰分析）をおこなうと，売上額と宣伝費，営業マンの数の間の関係式として次の回帰方程式が得られます。

$$y = 0.6785x + 0.6377u + 0.8739 \tag{2.1}$$

変量 x, u の係数の単位は売上額の単位千万円です。ちなみに（2.1）を使って A〜F の営業所の売上額を計算すると，実際の売上額とほぼ一致することが分かります（当てはまり具合の相当良い回帰式ということですね)。そこでこの式を使って G 営業所の売上額を見積ると

$$G \text{ 営業所の売上額（予測)} = 0.6785 \times 13 + 0.6377 \times 14 + 0.8739 = 18.6183 \text{ (千万円)}$$

と見込むことができます。いかがでしょうか。

2.1　重回帰分析

重回帰分析というのは，上の例でみたようにいくつかの原因（営業マンの数，宣伝費）と結果（売上高）を結ぶものということがいえます。結果の変量を**目的変量**と呼び，その原因をなす変量を**説明変量**といっています。要約すると

$$Y = a_1x_1 + a_2x_2 + a_2x_3 + \cdots + a_nx_n + a_0$$

という**重回帰方程式**を作り，この式を使って目的変量 y を予測したり，いろいろ分析する手法ということになります。尚，$a_1, a_2, a_3, \cdots, a_n$ を偏回帰係数といいます（説明変量が 1 つのときは**単回**

帰分析といい，これは既に学習しましたね)。この値は説明変量 $x_1, x_2, \cdots, x_n$ へのそれぞれの重みを意味し，その大きさが各説明変量の説明力の大きさを表します。もっと具体的に言うと，例えば 2 説明変量の場合を取り上げると，

$$Y = a_1x_1 + a_2x_2 + a_0 \tag{2.2}$$

上の重回帰式の偏回帰係数 a_1 は，説明変量 x_2 の影響を取り除いた目的変量 y と説明変量 x_1 との単回帰式の回帰係数に等しいということです。

さて，説明変量が 2 つの場合を取り上げて重回帰分析をみていくことにしましょう。重回帰方程を次式で表すします。

$$Y = ax + bu + c \quad (a, b, c \text{ は定数}) \tag{2.3}$$

問題は，偏回帰係数 a, b, c をどのようにして求めるかということですが，単回帰分析のところでやったように最小二乗法を使います。そこで，実測値 y_i と予測値 Y_i との残差を ε_i とおいて，

$$\varepsilon_i = y_i - Y_i = y_i - (ax_i + bu_i + c) \tag{2.4}$$

最小二乗法はこの残差 ε_i の 2 乗の総和 S を最小にするものでした。

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - ax_i - bu_i - c)^2 \tag{2.5}$$

S を最小にする条件は極値条件なので

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\displaystyle\sum_{i=1}^{n} x_i(y_i - ax_i - bu_i - c) = -2\sum_{i=1}^{n} x_i\varepsilon_i = 0 \\ \dfrac{\partial S}{\partial b} = -2\displaystyle\sum_{i=1}^{n} u_i(y_i - ax_i - bu_i - c) = -2\sum_{i=1}^{n} u_i\varepsilon_i = 0 \\ \dfrac{\partial S}{\partial c} = -2\displaystyle\sum_{i=1}^{n} (y_i - ax_i - bu_i - c) = -2\sum_{i=1}^{n} \varepsilon_i = 0 \end{cases} \tag{2.6}$$

となります。係数 c は各変量の平均で表すことができて

$$\begin{aligned} & \frac{\partial S}{\partial c} = -2\sum_{i=1}^{n} (y_i - ax_i - bu_i - c) = -2\sum_{i=1}^{n} \varepsilon_i = 0 \\ \therefore \quad & \frac{1}{n}\sum_{i=1}^{n} (y_i - ax_i - bu_i - c) = \bar{y} - a\bar{x} - b\bar{u} - c = 0 \\ \therefore \quad & c = \bar{y} - a\bar{x} - b\bar{u} \end{aligned} \tag{2.7}$$

これを（2.5）に入れると

$$S = \sum_{i=1}^{n} \{(y_i - \bar{y}) - a(x_i - \bar{x}) - b(u_i - \bar{u})\}^2 \tag{2.8}$$

これから

$$\begin{aligned} \frac{\partial S}{\partial a} &= -2\sum_{i=1}^{n} (x_i - \bar{x})\{(y_i - \bar{y}) - a(x_i - \bar{x}) - b(u_i - \bar{u})\} \\ &= 2\sum_{i=1}^{n} \{a(x_i - \bar{x})^2 + b(x_i - \bar{x})(u_i - \bar{u}) - (x_i - \bar{x})(y_i - \bar{y})\} \\ &= 2(as_x^2 + bs_{xu} - s_{xy}) = 0 \\ \therefore \quad as_x^2 + bs_{xu} &= s_{xy} \end{aligned} \tag{2.9}$$

が得られます。全く同様にして $\frac{\partial S}{\partial b} = 0$ より

$$as_{xu} + bs_u^2 = s_{uy} \tag{2.10}$$

したがって，偏回帰係数 a, b は次の連立方程式の解として求められます。

$$\begin{cases} s_x^2 a + s_{xu} b = s_{xy} \\ s_{xu} a + s_u^2 b = s_{uy} \end{cases} \tag{2.11}$$

ただし

$$\begin{aligned} s_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, & s_{xu} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(u_i - \bar{u}) \\ s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), & s_u^2 &= \sum_{i=1}^{n}\frac{1}{n-1}(u_i - \bar{u})^2 \\ s_{uy} &= \frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{u})(y_i - \bar{y}) \end{aligned} \tag{2.12}$$

連立方程式の問題は次の行列形式で書くと見通しがよく，解はクラメルの公式を使って容易に求めることができます。

$$\begin{pmatrix} s_x^2 & s_{xu} \\ s_{xu} & s_u^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} s_{xy} \\ s_{uy} \end{pmatrix} \tag{2.13}$$

ちなみに，左辺の行列

$$S = \begin{pmatrix} s_x^2 & s_{xu} \\ s_{xu} & s_u^2 \end{pmatrix} \tag{2.14}$$

を**分散・共分散行列**と呼んでいます。さて，クラメルの公式より，解は

$$\begin{cases} a = \dfrac{s_u^2 s_{xy} - s_{xu} s_{uy}}{s_x^2 s_u^2 - s_{xu}} \\ b = \dfrac{s_x^2 s_{uy} - s_{xu} s_{xy}}{s_x^2 s_u^2 - s_{xu}} \end{cases} \tag{2.15}$$

と得られ，あとは各分散をデータより計算すれば，a, b の値が求まります。

3 変量の場合への拡張は容易で，重回帰方程式を

$$Y = ax + bu + cv + d \tag{2.16}$$

とおくと，偏回帰係数 a, b, c は次の連立方程式の解となります。

$$\begin{pmatrix} s_x^2 & s_{xu} & s_{xv} \\ s_{xu} & s_u^2 & s_{uv} \\ s_{xv} & s_{uv} & s_v^2 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} s_{xy} \\ s_{uy} \\ s_{vy} \end{pmatrix} \tag{2.17}$$

定数項 d は回帰直線が分布の中心を通るということから

$$\bar{y} = a\bar{x} + b\bar{u} + c\bar{v} + d \longrightarrow d = \bar{y} - (a\bar{x} + b\bar{u} + c\bar{v})$$

として得られます。以上は理論的な話の筋道で，実際の計算は（手計算では大変なので：笑い）メモに書いたようにエクセルを使います。

さて，このセクションの最初で，ある販売会社の 6 営業所における宣伝費 (x)，営業マンの数 (u)，売上額 (y) のデータを重回帰分析して次の帰方程式を得ました。

$$Y = 0.6785x + 0.6377u + 0.8739$$

予測売上額を Y とすると，宣伝費を 100 万円上げると売上げは 0.68×1 千万円 $= 680$ 万円 上がり，営業マンの数を 1 人増やすと 0.64×1 千万円 $= 640$ 万円 上がることになります。つまり，宣伝費を 100 万円上げるのと営業マン 1 人を増やすのとは同じ効果が得られるということになります。このように，重回帰分析は与えられた資料から重要な結果を簡単に得ることができます。

＜メモ：エクセルによる重回帰分析＞
メニューバーの「ツール」→「分析ツール」→「回帰分析」と進んで「入力 Y 範囲（Y)」をマウスで選択します。今の場合，売上額 (y) の列の売上額 (y) の欄から F までを選択入力します。次に「入力 X 範囲（X)」は宣伝費 (x) と営業マンの数 (u) の 2 列で，その 2 列の欄から F までを選択入力します。次ぎに「ラベル」にチェックを入れ，出力オプションで「新規ブック（W)」をマークし，OK を押せば偏回帰係数が得られます。ただし各係数の値は次の表で与えられます。

	係数
切片	0.873889876
広告費	0.678507993
営業担当者数	0.637655417

$$Y = 0.6785\,x + 0.6377\,u + 0.8739$$

このように，重回帰分析は強力な武器ではありますが，次に述べる**多重共線性**（マルチコ）という問題があり，いつもうまく行くとは限りません。

2.1.1 多重共線性について

説明変量間で互いに強い相関があるとき，言い換えると，説明変量に互いに相関が強いものをいくつか選んでしまった場合，偏回帰係数が求められないということが起こります。仮に求まったとしてもその信頼性は低いものになります。これを**多重共線性**（「マルチコ」）といっています。その辺りの事情を簡単に見るために 2 説明変量の場合を考えて見ます。

（2.13）を

$$\boldsymbol{\alpha} = \begin{pmatrix} s_x^2 & s_{xu} \\ s_{xu} & s_u^2 \end{pmatrix}, \quad \boldsymbol{x} = \begin{pmatrix} a \\ b \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} s_{xy} \\ s_{uy} \end{pmatrix}$$

とおくと，解 $\boldsymbol{x}$ は

$$\boldsymbol{\alpha x} = \boldsymbol{\beta} \longrightarrow \boldsymbol{x} = \boldsymbol{\alpha}^{-1}\boldsymbol{\beta} \quad （ただし \boldsymbol{\alpha}^{-1} \neq 0） \quad より$$

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} s_x^2 & s_{xu} \\ s_{xu} & s_u^2 \end{pmatrix}^{-1} \begin{pmatrix} s_{xy} \\ s_{uy} \end{pmatrix} = \frac{1}{s_x^2 s_u^2 - s_{xu}^2} \begin{pmatrix} s_u^2 & -s_{xu} \\ -s_{xu} & s_x^2 \end{pmatrix} \begin{pmatrix} s_{xy} \\ s_{uy} \end{pmatrix}$$

$$= \frac{1}{s_x^2 s_u^2 - s_{xu}^2} \begin{pmatrix} s_u^2 s_{xy} - s_{xu} s_{uy} \\ -s_{xu} s_{xy} + s_x^2 s_{uy} \end{pmatrix}$$

と形式的に求まりますが，逆行列 $\boldsymbol{\alpha}^{-1}$ が存在しない場合には解 $\boldsymbol{x}$ は不定となり，方程式は解けないことになります。このような状況をに 多重共線性が存在する といいます。もう少し詳しくみると，

$$s_x^2 s_u^2 - s_{xu}^2 = 0 \quad \Longleftrightarrow \quad \frac{s_{xu}^2}{s_x^2 \cdot s_u^2} = 1 \quad \longrightarrow \quad r_{xu}^2 = 1 \quad \therefore\ r_{xu} = \pm 1$$

となって，x と u の相関係数が 1 または -1 のとき，つまり点 (x_i, u_i) のすべてが 共 通の直 線 上（共線）にある場合に多重共線性が存在する[8]ことになります。この問題は x と u が強い相関関係にありながらも，知らずして説明変量に組み込んだことが原因となっているので，多重共線性を避けるにはどちらか一方の説明変量を省略すればよいことになります。つまり，x_i あるいは u_i は一方

[8] 相関行列（R）の行列式の値が 0 に近いときは，多重共線性の状態にあるといいます。

が定まれば他方は直線関係から定まるので，どちらか一方の解釈のしやすい方を説明変量として採用しておけば良いということになります。説明変量の選択の基準は以下のようです。

(1) 説明変量間の相関が小さいもの

(2) 目的変量の予測に役立つもの

(3) 測定・管理のしやすいもの

さて， 多重共線性の具体的な例を見ていきましょう。ある町におけるパソコンの保有台数，世帯数，所得額のデータが下表のようになってたとします。

＜データ＞

	A	B	C	D
1		PC保有台数(y)	世帯数(x)	所得額(u)(億円)
2	A町	300	2200	198
3	B町	225	2000	200
4	C町	160	3000	240
5	D町	90	1300	91
6	E町	50	1500	75
7	F町	30	1000	55

＜相関係数＞

	PC保有台数	世帯数	所得額(億円)
PC保有台数	1		
世帯数	0.639770363	1	
所得額(億円)	0.819674256	0.9402682	1

＜重回帰分析結果＞

	係数
切片	82.1941736
世帯数	−0.165225
所得額(億円)	2.537031398

Y=−0.165252x＋2.5373u＋82.1942

回帰統計	
重相関 R	0.905434
重決定 R2	0.81981
補正 R2	0.699684
標準誤差	57.93773
観測数	6

相関係数をエクセルで計算してみましょう。「ツール」→「分析ツール」→「相関」を選び，入力範囲として“ $B$1:D7$ ”をマウスで選択します。「先頭行をラベルとして使用」をチェック，出力オプションで「新規ブック」をチェックすると＜相関係数＞のテーブルが出力されます。次に「分析ツール」→「回帰分析」を選択し，入力 Y 範囲として“ $B$1:B7$，入力 X 範囲として“ $C$1:D7$を入力,「ラベル」と「新規ブック」にチェックを入れ OK を押すと重回帰分析の結果が出力されます（表はその一部を抜き取ったものです）。重回帰分析の結果から重回帰方程式として

$$Y = -0.165252\,x + 2.5373\,u + 82.1942 \tag{2.18}$$

が得られます。この式から判断すると，所得額 1 億円当たりの PC 保有台数は −0.17 台 と所得額が多くなるにしたがって PC 保有台数は減少するという妙な結果になります。これが多重共線性（マルチコ）の引き起こす現象です。相関係数のテーブルより世帯数と所得額との相関係数は，0.9403 と大変強い相関を持っており，この 2 つの説明変量 x, u を回帰方程式の変数に入れたことがまずかったということになります。そこで，この問題を避けるために“ 説明変量相関で高い相関のあるものを探し，どちらかの変数を落とす ”という処方箋にしたがって，世帯数と所得額のいずれかの 1 つは説明変量として使用しないことにします。使用しない変数は，目的変数である「PC の保有台数」との 相関が低い 方の世帯数とします。そこで重回帰方程式を $Y = au + b$ として偏回帰係数を求めると $a = 1.104948, b = -15.6918$ が得られます。決定係数は $R^2 = 0.671866$ となるので，得られた回帰方程式

$$Y = 1.104948\,u - 15.6918$$

の精度はあまりよくないことに注意が必要です。

上の例で見てきたように，多重共線性は偏回帰係数の符号が相関係数の符号と一致しているかどうかを調べることで簡単に発見することができます。

＜多重共変性の見つけ方＞

PC の保有台数 $(Y) = a \times$ 世代数 $(x) + b \times$ 所得額 $(u) + c$

↕ ……→ （符号が一致しているか？）

Y と x の相関係数　Y と u の相関係数

＜メモ：説明変量の選び方＞
説明変量の選択方法として「変量減少法」「変量増加法」「総当り法」などがあり，その概要は次の通りです（詳しい内容は統計学の適当なテキストを参照ください）。

- **変量減少法：**説明変量として考えられるものすべてを取りあげます。それが N 個とすると，N 個の説明変量の寄与率 R_N^2（次のセクションを参照）を求め，次ぎに説明変量を 1 個減らした $N-1$ 個の説明変量の寄与率のうちで，R_N^2 との差が最小となる寄与率 R_{N-1}^2 を与える説明変量の組を残します。次ぎに R_{N-1}^2 との差が最小となる $N-2$ 個の説明変量の組を残します。以下，この操作を繰り返していきますが，あらかじめ決めた寄与率の値以下になったとき，あるいはあらかじめ決めた説明変量の個数になったときに操作を打ち切ります。
- **変量増加法：**採用する説明変量を十分に選んでおいて，それぞれ 1 個の説明変量の寄与率を計算し，その最大のものを残します。次ぎにその説明変量に残りの $N-1$ 個の説明変量をそれぞれ加えてみて，最大の寄与率を与える説明変量を残します。以下その操作を繰り返します。取り込むべき変量がなくなったときか，あらかじめ決めた寄与率の値以下になったときに打ち切ります。
- **総あたり法：**説明変量が N 個あるとすると，N 個すべての組み合わせ（2^N-1 個）について重回帰式を検討していく方法です。ただ，この方法は説明変量の個数が増えると作成する回帰式の数が膨大[9]なものになり，あまり実用的ではありません。

2.1.2 重相関係数と寄与率

重相関係数や**寄与率**についてはセクション 1.5.3 の決定係数のところで少し触れましたが，ここでもう一度取りあげておきます。重相関係数 R は実測値データと重回帰式から求めた予測値 Y との相関係数で，次式で定義されます。

$$\text{重相関係数 } R = r_{yY} = \frac{s_{yY}}{s_y \cdot s_Y} = \frac{\sum(y_i-\bar{y})\cdot(Y_i-\bar{y})}{\sqrt{\sum(y_i-\bar{y})^2}\cdot\sqrt{(Y_i-\bar{y})^2}} = \sqrt{\frac{\sum(Y_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}} = \sqrt{\frac{s_Y^2}{s_y^2}} \quad (2.19)$$

この式の右辺は次のようにして導出されます。

$$\begin{aligned}
s_{yY} &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})(Y_i-\bar{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-Y_i+Y_i-\bar{y})(Y_i-\bar{y}) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(\varepsilon_i+Y_i-\bar{y})(Y_i-\bar{y}) = \frac{1}{n-1}\left\{\sum_{i=1}^{n}\varepsilon_i(ax_i+bu_i+c-\bar{y})+\sum_{i=1}^{n}(Y_i-\bar{y})^2\right\} \\
&= \frac{1}{n-1}\left\{(c-\bar{y})\sum_{i=1}^{n}\varepsilon_i + a\sum_{i=1}^{n}\varepsilon_i x_i + b\sum_{i=1}^{n}\varepsilon_i u_i\right\} + \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{y})^2 \\
&= s_Y^2 \quad \text{(最後の式の右辺第 1 項は（2.6）よりゼロとなります。)} \\
\therefore\quad R &= \frac{s_{yY}}{s_y\cdot s_Y} = \frac{s_Y^2}{s_y\cdot s_Y} = \frac{s_Y}{s_y} = \sqrt{\frac{\sum(Y_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}} = \sqrt{\frac{s_Y^2}{s_y^2}}
\end{aligned}$$

重相関係数を 2 乗した R^2 を**寄与率**（決定係数）と呼び，$0 \leqq R^2 \leqq 1$ の値をとります。R^2 が 1 に近いほど重回帰式の精度が高い[10]ということになります。

2.2 おさらい

次のセクションの主成分分析にいく前に，具体例を元に今までのお話の総復習をやっておきましょう。材料工学分野に目を転じてセラミックスの焼結の話題を取り上げま

試料No	加圧条件			配向度
	温度(° C)	圧力(Mpa)	時間(Min)	%
1	1700	25	30	36
2	1800	15	25	39
3	1800	20	20	44
4	1850	20	30	44
5	1900	20	10	59
6	1930	20	10	51

(セラミックスの超精密加工：日刊工業新聞社　昭和57年)

[9] セクション 1.5.3 の《注意》を参照。

す[11]。セラミックスを高温･高圧化下の置くと内部の微結晶の結晶方位はある一定の方向を向くという配向現象を起こします。この結晶軸の並びを配向度といいますが，配向度が大きければ結晶軸の並びがそろった状態で，逆に小さければ結晶軸が無秩序になっている状態です。右の表は $\beta-$ アルミナ焼結体の鍛造条件と配向度についてのデータです。それでは早速やっていきましょう。

2.2.1 配向度を目的変量，温度と時間を説明変量にとり，重回帰方程式を求める

試料 No3〜6 では圧力がすべて同じで，温度と時間の 2 つのパラメータで配向度が変化していることがわかります。そこで目的変量に配向度をとり，説明変量に温度と時間を選択して，配向度に対して温度と時間が及ぼす影響をエクセルを使って重回帰分析することにします。求める重回帰方程式を

$$\begin{aligned} \text{配向度} &= a_1 \times \text{温度} + a_2 \times \text{時間} + a_0 \\ Y \quad &= a_1\, x_1 + a_2\, x_2 + a_0 \end{aligned} \tag{2.20}$$

としておきます。まず次ページの表の上段右にあるような表を作ります。次ぎにエクセルの「ツール」→「分析ツール」→「回帰分析」→入力 Y 範囲 (Y):で`$C$1:$C$7`，入力 X 範囲 (X):で`$A$1:$B$7`を入力し，ラベルをチェック，出力オプションでは新規ブックをチェック→ OK という手順を踏むと以下の表が出力されます。

[11] 有馬哲，石村貞夫（著）「多変量解析のはなし」東京図書を参考にしました。

概要

回帰統計	
重相関 R	0.911803
重決定 R2	0.831385
補正 R2	0.718976
標準誤差	4.43211
観測数	6

	A	B	C
1	温度(° C)	時間(Min)	%
2	1700	30	36
3	1800	25	39
4	1800	20	44
5	1850	30	44
6	1900	10	59
7	1930	10	51

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	2	290.5692	145.2846	7.396027016	0.069238
残差	3	58.9308	19.6436		
合計	5	349.5			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	-28.411	76.03733	-0.37365	0.733510756	-270.396	213.5737	-270.396	213.5737
温度(° C)	0.045734	0.038401	1.19094	0.319317523	-0.07648	0.167944	-0.07648	0.167944
時間(Min)	-0.46951	0.345168	-1.36024	0.266942618	-1.56799	0.628966	-1.56799	0.628966

残差出力

観測値	予測値 : %	残差
1	35.25078	0.74922
2	42.1717	-3.1717
3	44.51925	-0.51925
4	42.11082	1.889178
5	53.78772	5.212279
6	55.15973	-4.15973

確率

百分位数	%
8.333333333	36
25	39
41.66666667	44
58.33333333	44
75	51
91.66666667	59

これから偏回帰係数を求めると $a_1 = 0.045734$, $a_2 = -0.46951$, $a_0 = -28.411$ となるので，重回帰方程式は

$$Y = 0.045734\, x_1 - 0.46951\, x_2 - 28.411 \tag{2.21}$$

と求まります。ついでに分散・共分散行列を求めておきます。**分散・共分散行列**は「セクション 2.1 重回帰分析」の（2.14）で紹介しましたが，繰り返すと例えば 3 変量 x, y, z の場合

$$\begin{aligned} S &= \begin{pmatrix} x\text{の分散} & x\text{と}y\text{の共分散} & x\text{と}z\text{の共分散} \\ x\text{と}y\text{の共分散} & y\text{の分散} & y\text{と}z\text{の共分散} \\ x\text{と}z\text{の分散} & y\text{と}z\text{の共分散} & z\text{の分散} \end{pmatrix} \\ &= \begin{pmatrix} s_x^2 & s_{xy} & s_{xz} \\ s_{xy} & s_y^2 & s_{yz} \\ s_{xz} & s_{yz} & s_z^2 \end{pmatrix} \end{aligned} \tag{2.22}$$

$$s_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2, s_{xy} = \frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y}), s_{xz} = \frac{1}{n-1}\sum (x_i - \bar{x})(z_i - \bar{z})$$
$$s_y^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2, s_{yz} = \frac{1}{n-1}\sum (y_i - \bar{y})(z_i - \bar{z})$$
$$s_z^2 = \frac{1}{n-1}\sum (z_i - \bar{z})^2$$

というものです。今の例に当てはめると

$$\begin{cases} s_x^2 \to s_1^2\text{: 温度の分散}, & s_{xy} \to s_{12}\text{: 温度と時間の共分散}, & s_{xz} \to s_{1y}\text{: 温度と配向度の共分散} \\ s_y^2 \to s_2^2\text{: 時間の分散}, & s_{yz} \to s_{2y}\text{: 時間と配向度の共分散} & \\ s_z^2 \to s_y^2\text{: 配向度の分散} & & \end{cases}$$

尚，誤解はないと思いますが，下付きの数字は変量 x_1, x_2 を表すものです，念のため。分散・共分散を手計算でやるのはしんどいのでエクセルの関数を使いますでやります。分散は，メニューの「挿入」→「関数」から VAR(数値 1, 数値 2, ⋯) 関数を選び，対象となる数値を入れれば OK。共分散も同じく「関数」から COVAR(配列 1, 配列 2) 関数で計算できますが，この関数は分母が n であるため，COVAR(配列 1, 配列 2) $\times \frac{n}{n-1}$ として補正しておく必要があります。そうして分散・共分散行列は次のようになります。

$$S = \begin{pmatrix} s_1^2 & s_{12} & s_{1y} \\ s_{12} & s_2^2 & s_{2y} \\ s_{1y} & s_{2y} & s_y^2 \end{pmatrix} = \begin{pmatrix} 6800 & -590 & 588 \\ -590 & 84.2 & -66.5 \\ 588 & -66.5 & 69.9 \end{pmatrix}$$

また，重回帰式（2.20）の偏回帰係数 a_1, a_2 と分散・共分散行列は次式で表されることは既に（2.14）で見てきた通りです。

$$\begin{pmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} s_{1y} \\ s_{2y} \end{pmatrix}$$

2.2.2 重回帰式の当てはまり具合を調べる

式（2.21）で得られた重回帰式の当てはまり具合を調べて見ましょう。右の表は実測値と予測値を比較したものです。セクション 1.5.3 の式（1.20）で示したように

試料No	実測値y	予測値Y	残差(ε)
1	36	35.27	0.73
2	39	42.19	-3.19
3	44	44.53	-0.53
4	44	42.13	1.87
5	59	53.80	5.20
6	51	55.17	-4.17
平均	45.5	45.5	0.0
分散	69.9	58.0	11.8

実測値の分散＝予測値の分散＋残差の分散

実測値の分散＝予測値の分散＋残差の分散

という関係式が成立していることが分かります。実測値の分散はデータが与えられたときにある決まった値となるので，残差の分散が小さいということがとりもなおさず重回帰式の当てはまり具合がよいという結論になります。そこで，当てはまり具合の良さを表す指標として**決定係数**あるいは**寄与率**を次のように定義しました。

$$R^2 = \frac{\text{(目的変量の) 予測値の分散}}{\text{(目的変量の) 実測値の分散}}, \quad (0 \leqq R^2 \leqq 1) \tag{2.23}$$

いまの場合は

$$R^2 = \frac{58.0}{69.9} = 0.830$$

となって，当てはまり具合はよろしいということになります。また，決定係数（寄与率）の平方根 R を**重相関係数**といいました。重相関係数は（2.19）で

$$R = r_{yY} = \frac{\sum (y_i - \bar{y}) \cdot (Y_i - \bar{y})}{\sqrt{\sum (y_i - \bar{y})^2} \cdot \sqrt{(Y_i - \bar{y})^2}}$$

と定義したように，実測値と予測値との間の相関係数を意味しています。

決定係数や重相関係数が抱える問題として，説明変量の数を増やすと大きくなり，より 1 に近付くという性質があります。したがって，決定係数や重相関係数が大きいから当てはまり具合が良いとは単純にはいえません。この辺りの事情を以下に見ていくことにしましょう。

2.2.3　説明変量を増やしてみると

配向度に影響をあまり与えないと考えられる圧力を説明変量 x_3 として加えてみます。重回帰方程式は次のようになります。

$$Y = a_1x_1 + a_2x_2 + a_3x_3 + a_0 \tag{2.24}$$

上でやったようにエクセルで計算すると

概要

回帰統計	
重相関 R	0.926538
重決定 R2	0.858473
補正 R2	0.646182
標準誤差	4.973115
観測数	6

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	3	300.0363	100.0120835	4.043853707	0.204592
残差	2	49.46375	24.73187477		
合計	5	349.5			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	-60.7349	100.0442	-0.6070808	0.605537715	-491.19	369.7204	-491.19	369.7204
温度(° C)	0.057506	0.047103	1.220859056	0.346535642	-0.14516	0.260175	-0.14516	0.260175
圧力(Mpa)	0.4829	0.78051	0.618698204	0.599192166	-2.87536	3.841163	-2.87536	3.841163
時間(Min)	-0.41567	0.396956	-1.04714877	0.404925411	-2.12364	1.292292	-2.12364	1.292292

という表が出力されます。これから重回帰方程式と決定係数 R^2，重相関係数 R は

$$\begin{cases} \text{重回帰方程式：} & Y = 0.057506x_1 - 0.41567x_2 + 0.4829x_3 - 60.7349 \\ \text{決定係数 (寄与率)：} & R^2 = 0.858473 \\ \text{重相関係数：} & R = 0.926538 \end{cases} \tag{2.25}$$

と求まります。さて，ここで温度，時間，圧力と 3 つの説明変量をとったときの決定係数 R^2 は，温度と時間の 2 つの説明変量をとった先ほどの重回帰式の決定係数 $R^2 = 0.831385$ より大きな値となりました。そこで，この回帰式は先ほどのものより，より当てはまり具合の良い式だと喜びたいですが，現実はそう甘くない。。。ではどう甘くないか，それは重回帰の検定を行うと 3 説明変量の重回帰方程式は予測に役立たないという結果がでてくるのです！

2.2.4　重回帰の検定

●準備

重回帰分析の場合，結果の正当性は決定係数 R^2 で判断されます。R^2 が 1 に近ければよい結果ということになりますが，それを単純に鵜呑みするのは危険ということですね。上の例にみたように，目的変量にほとんど影響を及ぼさないと考えられる説明変量を加えると R^2 は大きくなります。ということで R^2 は現象を本当に反映しているのか，それを検定していく必要があります。式（1.20）より

$$\text{目的変量の実測値の分散 } (s_y^2) = \text{予測値の分散 } (s_Y^2) + \text{残差の分散 } (s_\varepsilon^2)$$

また,（1.21）より重相関係数は

$$R^2 = \frac{\text{目的変量の予測値の分散}}{\text{目的変量の実測値の分散}} = \frac{s_Y^2}{s_y^2}$$

上の 2 式より

$$R^2 = \frac{\text{予測値の偏差平方和}}{\text{予測値の偏差平方和＋残差の偏差平方和}} = \frac{1}{1 + \dfrac{\text{残差の偏差平方和}}{\text{予測値の偏差平方和}}}$$

となるので，R^2 が大きいことは次の比 F_R

$$F_R = \frac{\text{予測値の偏差平方和}}{\text{残差の偏差平方和}}$$

が大きいことと同値なので，R^2 が大きいことを検定するかわりに F_R が大きいことを検定すればよいことになります。ところで，

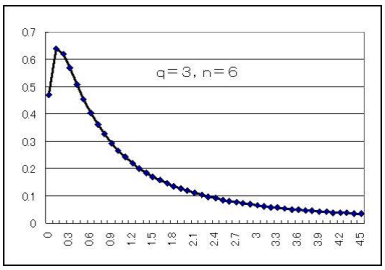
$$F_0 = \frac{V_Y}{V_\epsilon} = \frac{\text{予測値の偏差平方和／説明変数の数}}{\text{残差平方和／ (サンプル数 − 説明変数の数 − 1)}} = \frac{\text{予測値の偏差平方和／ } q}{\text{残差平方和／ } (n-q-1)}$$

という値を導入すると

$$F_0 = \frac{V_Y}{V_\varepsilon} = \frac{n-q-1}{q} F_R \tag{2.26}$$

となるので，F_R のかわりに F_0 が大きいことを検定すればよい。F_0 を**分散比**と呼んでいます。

なぜこのようなことをするのかというと，F_0 は自由度（$q, n-q-1$）の $F-$ 分布に従うからなんです[12]（F 分布の具体的な形は右図参照)。F_0 を手計算するのは面倒ですが，先ほどのエクセルの表での「観測された分散比」の欄が F_0 を与えてくれています。



●重回帰の検定

検定を行うにあたって次の仮説を立てます

《仮説 H_0》：求めた重回帰式は目的変量の予測に役に立たない。

この仮説に対して，分散比 F_0 が

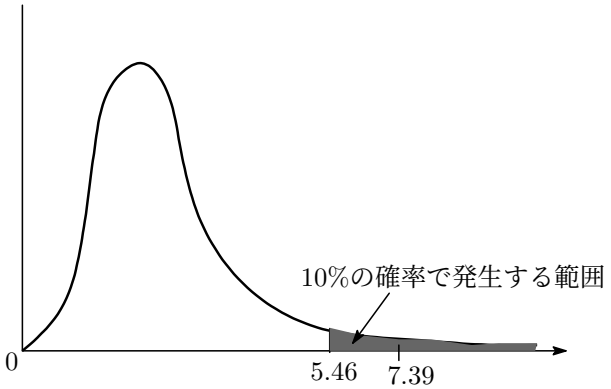
$$F_0 \geqq F(q, n-q-1; \alpha) \tag{2.27}$$

となるなら，仮説 H_0 は危険率（あるいは有意水準）$\alpha\%$ で棄却される，というのが F 検定と呼ばれるものです。$F(q, n-q-1; \alpha)$。それでは具体的に検定に入ります。危険率を 10%とします。

A．重回帰式（2.21) の検定 説明変量が「温度」と「時間」の 2 つですから $q = 2$，$n = 6$ なので，数表より

$$F(2, 4; 0.1) = 5.4624$$

と求まります。F 分布に従う現象において $x \geqq 5.46$ となる確率は 10%であることを示し，一方，エクセルの表から分散比を求めると $F_0 = 7.3960$ なので（2.27）が成り立ち，仮説 H_0 は 10%の危険率で棄却されることになります。ということで重回帰式（2.21) の正しさが検定されました。



[12] 残差 ε_i が互いに独立で，正規分布に従うとの仮定があります。

B．重回帰式（2.25) の検定 説明変量が「温度」「時間」「圧力」の 3 つですから $q = 3$，$n = 6$ なので，数表より

$$F(3, 2; 0.1) = 9.1618$$

一方，エクセルの表から分散比を求めると $F_0 = 4.043$ なので（2.27）が成り立たず,

$$F_0 = 4.043 < 9.1618 = F(3, 2, 0.1)$$

この結果，重回帰式（2.25) は予測に役立つとはいえないということになります。

2.2.5　偏回帰係数の意味するところ

さて，おさらいの最後として偏回帰係数の意味するところを整理しておきます。セクション 2.1 の重回帰分析のところで，2 説明変量の重回帰式

$$Y = a_1 x_1 + a_2 x_2 + a_0$$

の偏回帰係数 a_1 は，説明変量 x_2 の影響を取り除いた目的変量 y と説明変量 x_1 との単回帰式の回帰係数に等しいということを述べました。このことをもう少し詳しく調べておきましょう。

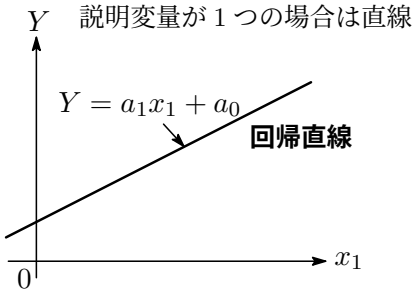
「温度」を説明変量にとった場合の配向度の単回帰式を求めると（エクセルを使うと容易ですね）

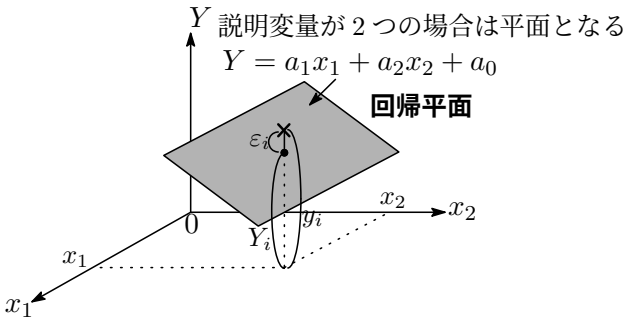
$$Y = 0.086471\, x_1 - 112.741 \tag{2.28}$$

となります。一方,「温度」と「時間」を説明変量にとった場合の重回帰方程式は次式で与えられました（(2.21）参照)。

$$Y = 0.045734\, x_1 - 0.46951\, x_2 - 28.411 \tag{2.29}$$

(2.28) は 2 次元平面で配向度 Y を温度 x_1 を従属変数とした直線で捉えおり, (2.29）は，3 次元空間で配向度 Y を温度 x_1，圧力 x_2 の 2 つを従属変数とする平面で捉えています。





問題は（2.28）が（2.29）で $x_2 = 0$ とした場合（$Y - x_1$ 平面へ射影）の直線の式にならないという点です。なぜそうならないのかということを次ぎに説明していきます。結論から先に言うと，単回帰方程式（2.28）の偏回帰係数 $a_1(= 0.086471)$ は時間の説明変量 x_2 の影響が混ざっており，x_2 の影響を取り除いた偏回帰係数が重回帰方程式（2.29）の偏回帰係数 $a_1(= 0.045734)$ であるということです。なぜ混ざり合うのか，x_1 と x_2 の相関係数を調べてみると（エクセルで簡単に計算できますね）相関係数 $r_{12} = -0.7799$ が得られます。つまり，温度と時間の間にはかなり大きな負の相関があるわけで，これが混ざり合いの要因となっています。

そこで時間の温度に与える影響を調べるために，温度 x_1 と時間 x_2 の単回帰方程式を求めると

$$X_1 = -7.0099 x_2 + 1976.04$$

試料No	温度x1	時間x2	予測値X1	残差 x_1−X_1=p
1	1700	30	1765.743	−65.743
2	1800	25	1800.793	−0.793
3	1800	20	1835.842	−35.842
4	1850	30	1765.743	84.257
5	1900	10	1905.941	−5.941
6	1930	10	1905.941	24.059

となります。右の表で，残差 p は温度から時間の影響を取り除いた量となります。次ぎに，時間は配向度にも影響を与えているので，配向度から時間の影響を取り除くために上と同じ様に単回帰の方程式を求めておくと

$$Y = -0.7901x_2 + 61.9604$$

が得られます。残差 q は配向度から時間の影響を取り除いた量です。

ということで，q と p の関係は時間の影響を取り除いた配向度と温度だけの関係となります。そこで，時間の影響を取り除いた配向度 (Y') に対する時間の影響を取り除いた圧力 (x_1') の関係の単回帰方程式を求めると

$$Y' = 0.045734x_1' + 3.57 \times 10^{-5}$$

試料No	配向度y	時間x2	予測値Y	残差 y−Y=q
1	36	30	38.257	−2.257
2	39	25	42.208	−3.208
3	44	20	46.158	−2.158
4	44	30	38.257	5.743
5	59	10	54.059	4.941
6	51	10	54.059	−3.059

となり，得られた回帰係数 0.045734 は 重回帰方程式の x_1 の偏回帰係数と一致 することが分かります。

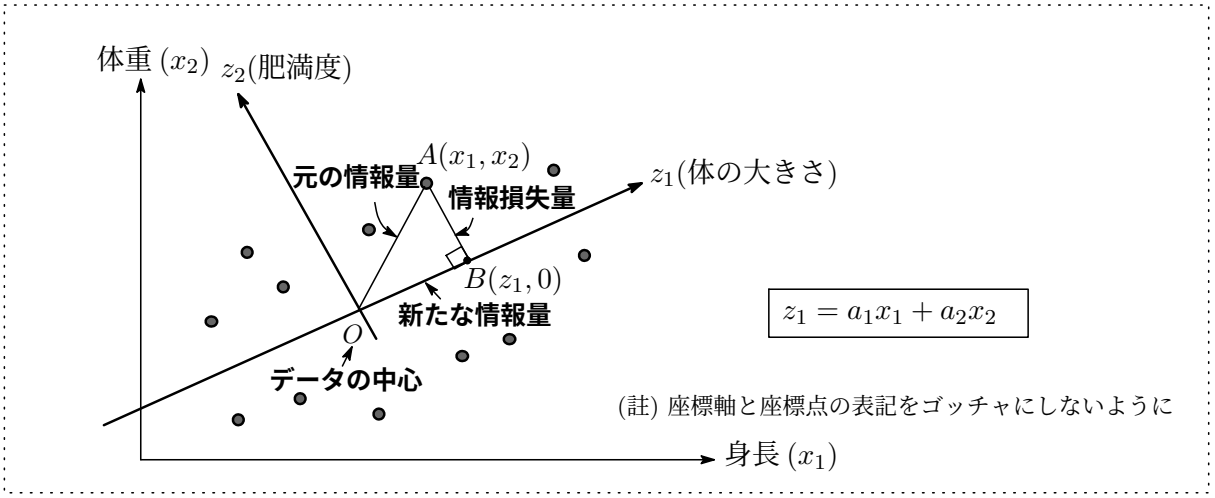
3 主成分分析

3.1 主成分分析のコンセプト

重回帰分析では目的変量に対する沢山の説明変量 $x_1, x_2, \cdots, x_p$ をそれぞれ独立に取り扱いました。 主成分分析は，以下のように言われる分析手法です。

- 相関関係の認められる p 個の変量の値を，少数個の合成変量で表すデータ処理法である。
- ある問題に対していくつかの要因が考えられるとき，それらの要因を一つ一つ独立に扱うのではなくて総合的に取り扱う手法である。
- 沢山の変量からなるデータをできるだけ情報を減らすことなくデータの縮約を行い，そのデータの背後にある構造を探り出す手法である。

これらの説明で成る程と頷ければそれに越したことはないですが，小生を含め大抵の方はピンとこないだろうと思います。そこで具体的な事例を取り上げて主成分分析のコンセプトに触れておくことにします。



ある学校の生徒の身体検査で身長と体重の測定結果が上図のようになったとします。この図から体重と身長は右肩上がりの正の相関関係を示していることがわかります。ところで，図に示すようにデータの中心 O を通る z_1 軸を新たに考えると，z_1 軸上の点は身長が大きくなるほど体重も大きくなるという“ 体の大きさ ”を表すと考えることができますね。次ぎに z_1 軸と直交し，データの中心 O を通る z_2 軸を考えると，この軸上の点は“ 肥満の程度 ”を表すと見做すことができます。身長や体重といった変量だけで体型を把握するのは難しいですが，身長と体重を取り混ぜた“ 体の大きさ ”や“ 肥満の程度 ”という新たな変量を導入することで，総合的な体のイメージを掴むことができるようになります。主成分分析の主成分とは総合的な特性（今の場合は体型）のことで，“ 体の大きさ ”を表す z_1 を**第 1 主成分**，“ 肥満の程度 ”を表す z_2 を**第 2 主成分**と呼んでいます。

さて，新たなキーワードを導入しましたが，変量の数は何も変わっていなじゃないかと疑問を持たれたと思います。そこで，個々人の体型は“ 体の大きさ ”z_1 でほぼ表現できるという点に注目すれば 2 種類の変量で表されたデータは 1 種類の新たな変量で表現できることになりますね。たしかに身体検査の結果を“ 体の大きさ ”で代表させた場合“ 肥満の程度 ”についての情報は失われますが，この情報の損失量を最小化すればよいわけです。情報の損失量は，z_1z_2 平面で z_1 軸へ降ろした垂線の長さ $\overline{AB}$ で与えられ，新たな情報量は線分 $\overline{OB}$ で与えられるので，線分 $\overline{OA}$ を元の情報量とすると

$$\overline{OA}^2 = \overline{OB}^2 + \overline{AB}^2 \tag{3.1}$$

という関係が成立します。この関係式は得られたすべてのデータについて成り立つので，データ全体については

$$\text{「元の情報量の 2 乗和」} = \text{「新たな情報量の 2 乗和」} + \text{「情報損失量の 2 乗和」} \tag{3.2}$$

という関係が得られます。左辺はデータが与えられたときに決まる定数ですから「情報損失量の 2 乗和」を最小にすることは，とりもなおさず「新たな情報量の 2 乗和」を最大にすることになります。「新たな情報量の 2 乗和」を最大にするためには，図から分かるように分散が最大となる方向に z_1 軸を設定すれば良いということになりますが[13]，この点はまた後ほど詳しく見ていくことにして，新たな情報量 z_1 の絶対値 $|\, z_1 \,|$ を**主成分得点**と呼んでいる[14]ことを付け加えておきます。

以上，2 変量の場合について見てきましたが，主成分分析というのは，一般に p 個の変量 $x_i\,(i = 1, 2, \cdots p)$ の持つ情報を情報の損失を最小限に抑え，互いに独立な $m\,(< p)$ 個の総合的指標 $z_i\,(i = 1, 2, \cdots, m)$ に変換する手法のことで，新たな変量 z_i は次のように x_i の線形和で表わされます。

$$\begin{cases} z_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ z_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\ \qquad\qquad \vdots \\ z_m = a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mp}x_p \end{cases} \tag{3.3}$$

$z_1, z_2, \cdots, z_m$ をそれぞれ第 1 主成分，第 2 主成分，$\cdots$, 第 m 主成分といいます。

【※】大事な注意：イメージの掴みやすい例として身長と体重をあげましたが，この 2 つの変量は単位が異なるという点に注意ください。例えば身長は 1.7m と 170cm は同じ意味ですが 1.7m を採用した場合と 170cm を採用した場合の計算結果は当然異なってきます。このような不具合をなくすために単位の異なるデータの場合はセクション 1.4 でやったようにデータを標準化して単位を無次元化しておく必要があります。このケースはセクション 3.2.2 で取り扱います。

[13] 相関が小さいデータの場合は，データがあらゆる方向に均等に分布するので z_1 軸をどの方向にとっても分散を最大にする軸はありません。このようなデータの場合には主成分分析はうまくいきません。

[14] 負の情報量は意味不明となるので。

3.2 主成分の求め方

3.2.1 下準備

2 変量 x_1, x_2 の場合を考えてみます。第 1 主成分 z_1 は

$$z_1 = a_1 x_1 + a_2 x_2 \tag{3.4}$$

という一次式で表されます。情報損失量を最小化する係数 a_1, a_2 を求めるのが主成分の求め方となります。そこで係数 a_1, a_2 の幾何学的な意味を調べてみると，これは下図に示すように z_1 軸の傾きと関係していることが分かります。また，係数 a_1, a_2 の間には次の条件が付くことに留意ください。

$$a_1^2 + a_2^2 = 1 \tag{3.5}$$

したがって，z_1 軸上の点は x_1, x_2 平面で $\tan\theta$ の傾きを持つ直線上の点となるので

$$x_2 = \frac{a_2}{a_1} x_1 + c \longrightarrow a_2 x_1 - a_1 x_2 + a_0 = 0\,, \quad (a_0 = a_1 c) \tag{3.6}$$

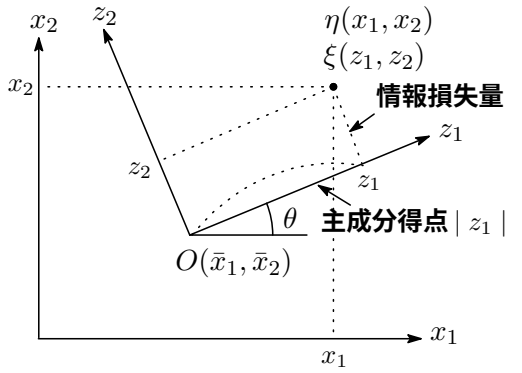
を満たします。次ぎに，情報損失量は図の点 $\eta(x_1, x_2)$ から z_1 軸に降ろした垂線の長さで，公式[15]と (3.5) より

$$\frac{\mid a_2 p_1 - a_1 q_1 + a_0 \mid}{\sqrt{a_2^2 + a_1^2}} = \mid a_2 p_1 - a_1 q_1 + a_0 \mid \tag{3.7}$$

で与えられます。なお，絶対値がついていると扱いにくいので 2 乗しておき，すべてのデータについてこの“ **情報損失量の平方の総和**”を $U(a_2, a_1, a_0)$ とすると，求める係数 a_1, a_2 は

$$\begin{cases} \text{条件：} a_2^2 + a_1^2 = 1 \quad \text{の下で} \\ U(a_2, a_1, a_0) \text{ の最小値を与える } a_1, a_2 \end{cases} \tag{3.8}$$

ということになります。この問題は変分法でお馴染みの**束縛条件付の極値問題**で，**ラグランジュの未定係数法**[16]を使って解くことができますが，このことはまた後ほど触れます。



$$\begin{aligned} z_1 &= (x_1 - \bar{x}_1)\cos\theta + (x_2 - \bar{x}_2) sin\theta \\ &= x_1 \cos\theta + x_2 \sin\theta - (\bar{x}_1 \cos\theta + \bar{x_2} \sin\theta) \\ &= a_1 x_1 + a_2 x_2 + a_0 \end{aligned}$$

ただし $a_1^2 + a_2^2 = 1$

$$\left(\tan\theta = \frac{a_2}{a_1}\right), \quad a_0 = -(\bar{x}_1 \cos\theta + \bar{x_2} \sin\theta)$$

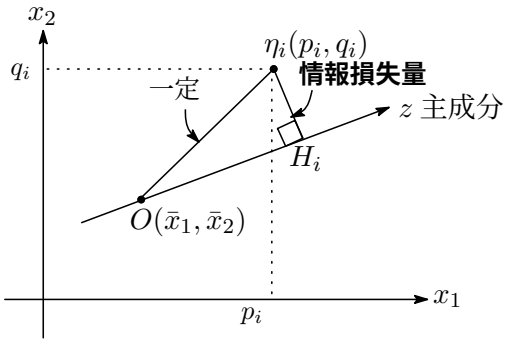
以上の話は大事なポイントなので冗長を省みず（笑い）復習をかねてもう少し詳しく見ていくことにします。下図を見ていただくとして，次の 3 平方の定理が成立します。

$$\overline{O\eta_i}^2 = \overline{OH_i}^2 + \overline{\eta_i H_i}^2 \tag{3.9}$$

---

[15] 点 (x_0, y_0) から直線 $ax + by + c = 0$ へ降ろした垂線の長さは $\frac{|ax_0 + by_0 + c|}{\sqrt{a^2 + b^2}}$

[16] ラグランジュの未定乗数法ともいわれます。



$$\overline{O\eta_1}^2 = (p_i - \bar{x}_1)^2 + (q_i - \bar{x}_2)^2$$
$$\overline{\eta_i H_i}^2 = (a_2 p_i - a_1 q_i + a_0)^2$$
$$= (a_2 p_i - a_1 q_i - a_2 \bar{x}_1 + a_1 \bar{x}_2)^2$$
$$= \{(p_i - \bar{x}_1)a_2 - (q_i - \bar{x}_2)a_1\}^2$$

主成分得点 $\overline{OH_i}$ の意味合いを調べるために $\overline{OH_i}^2$ を展開すると

$$\begin{aligned}
\overline{OH_i}^2 &= \overline{O\eta_i}^2 - \overline{\eta_i H_i}^2 \\
&= (p_i - \bar{x}_1)^2 + (q_i - \bar{x}_2)^2 - (p_i - \bar{x}_1)^2 a_2^2 - (q_i - \bar{x}_2)^2 a_1^2 + 2(p_i - \bar{x}_1)(q_i - \bar{x}_2)a_1 a_2 \\
&= (p_i - \bar{x}_1)^2(1 - a_2^2) + (q_i - \bar{x}_2)^2(1 - a_1^2) + 2(p_i - \bar{x}_1)(q_i - \bar{x}_2)a_1 a_2 \\
&= (p_i - \bar{x}_1)^2 a_1^2 + (q_i - \bar{x}_2)^2 a_2^2 + 2(p_i - \bar{x}_1)(q_i - \bar{x}_2)a_1 a_2 \quad (a_1^2 + a_2^2 = 1 \text{ を使った}) \\
&= \{(p_i - \bar{x}_1)a_1 + (q_i - \bar{x}_2)a_2\}^2 \\
&= \{(p_i a_1 + q_i a_2) - (\bar{x}_1 a_1 + \bar{x}_2 a_2)\}^2 = \{(\eta_i \text{の主成分の値}) - (\text{主成分の値の平均})\}^2
\end{aligned}$$

となり，これは主成分の分散を意味しますね。いまデータが n 個あるとして（3.9）の総和をとり，両辺を $n-1$ で割ると

$$\begin{aligned}
\frac{1}{n-1}\sum_{i=1}^{n}\overline{O\eta_i}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(p_i - \bar{x}_1)^2 + \frac{1}{n-1}\sum_{i=1}^{n}(q_i - \bar{x}_2)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\overline{OH_i}^2 + \frac{1}{n-1}\sum_{i=1}^{n}\overline{\eta_i H_i}^2
\end{aligned}$$

となり，まとめると

$$\textbf{説明変量 } x_1, x_2 \textbf{の分散の和} = \textbf{主成分の分散} + \textbf{情報損失量の 2 乗和} \tag{3.10}$$

ということになります。左辺はデータが与えられたとき一定の値をとるので，情報損失量の 2 乗和を最小にするということは 主成分の分散を最大にする ことになるわけです。したがって係数 a_1, a_2 は次の 2 通りの方法でそれぞれ求めることができます。$a_1^2 + a_2^2 = 1$ を条件として

$$\begin{cases} \text{主成分の分散を最大にする} \\ \text{情報損失量の 2 乗和を最小にする。} \end{cases} \tag{3.11}$$

以下，2 通りの方法でやってみます。そしてその結果は一致することを確認することにします。

3.2.2 主成分の分散を最大にする方法

主成分の分散を V と書くと

$$\begin{aligned} V &= \frac{1}{n-1}\sum_{i=1}^{n}\overline{OH_i}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\{(p_i a_1 + q_i a_2) - (\bar{x}_1 a_1 + \bar{x}_2 a_2)\}^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\{(p_i - \bar{x}_1)a_1 + (q_i - \bar{x}_2)a_2\}^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\{(p_i - \bar{x}_1)^2 a_1^2 + (q_i - \bar{x}_2)^2 a_2^2 + 2(p_i - \bar{x}_1)(q_i - \bar{x}_2)a_1 a_2\} \\ &= s_1^2 a_1^2 + s_2^2 a_2^2 + 2s_{12}a_1 a_2 \end{aligned} \tag{3.12}$$

ただし，

$$\begin{cases} s_1^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(p_i - \bar{x}_1)^2 \\ s_2^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(q_i - \bar{x}_2)^2 \\ s_{12} = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(p_i - \bar{x}_1)(q_i - \bar{x}_2) \end{cases} \tag{3.13}$$

で，主成分の分散 V は変量 x_1 の分散の項と x_2 の分散の項，さらに x_1と x_2 の共分散の項の足し算となります。ところで (3.12) は次ぎように行列形式で表すことができます。

$$V = \begin{pmatrix} a_1 & a_2 \end{pmatrix}\begin{pmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \tag{3.14}$$

さて，求める係数 a_1, a_2 は $a_1^2 + a_2^2 = 1$ という条件のもとで分散 V を最大にするものですが，このような条件付極値問題には**ラグランジュの未定乗数法**が使われます（←詳しいことは本稿末にメモとしてまとめておきましたので，ザット目を通しておいてください）。ラグランジュの未定乗数法を適用すると，新たな関数 G を導入し

$$G = s_1^2 a_1^2 + s_2^2 a_2^2 + 2s_{12}a_1 a_2 - \lambda(a_1^2 + a_2^2 - 1)$$

とおいて，この G の極値を与える a_1, a_2 は G をそれぞれの変数で偏微分したものが 0 となりますから

$$\begin{cases} \dfrac{\partial G}{\partial a_1} = 2s_1^2 a_1 + 2s_{12}a_2 - 2\lambda a_1 = 0 \Longrightarrow s_1^2 a_1 + s_{12}a_2 = \lambda a_1 \\ \dfrac{\partial G}{\partial a_2} = 2s_2^2 a_2 + 2s_{12}a_1 - 2\lambda a_2 = 0 \Longrightarrow s_{12}a_1 + s_2^2 a_2 = \lambda a_2 \end{cases}$$

これを行列形式で書き直すと

$$\begin{pmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \tag{3.15}$$

となります。左辺先頭の行列は**分散・共分散行列**ですね。求める係数 a_1, a_2 はこの行列方程式の解ですが，これは線形代数でお馴染みの列ベクトル $\boldsymbol{a}$ を固有ベクトル，λ を固有値とする固有値問題となります。

さて，(3.15) を (3.14) に入れると

$$\begin{aligned} V &= \begin{pmatrix} a_1 & a_2 \end{pmatrix}\begin{pmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} a_1 & a_2 \end{pmatrix}\lambda\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \\ &= \lambda(a_1^2 + a_2^2) = \lambda \quad (\because a_1^2 + a_2^2 = 1) \end{aligned} \tag{3.16}$$

となり，**固有値 λ は主成分の分散に等しい**という重要な結果がでてきます。この結果は，第 1 主成分の係数を決めるときの指針となりますので，頭に入れておいてください。

次ぎにデータを標準化すると，分散・共分散行列は**相関行列**になることを示しておきます。サフィックスのややこしさを避けるために 2 つの変数を x と y にし，標準化した変数を X, Y とすると

$$X_i = \frac{x_i - \bar{x}}{s_x}, \quad Y_i = \frac{y_i - \bar{x}}{s_y} \quad (s_x, s_y\text{: 標準偏差})$$

標準化した変数の平均は 0 になる[17]ので，標準化した変数の共分散 s_{XY} は

$$\begin{aligned} s_{XY} &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n-1}\sum_{i=1}^{n} X_i Y_i \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{s_x}\cdot\frac{y_i - \bar{y}}{s_y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \times \frac{1}{s_x s_y} = \frac{s_{xy}}{s_x s_y} \\ &= r_{xy} \\ s_{XX} &= \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \frac{1}{s_x^2} = s_x^2/s_x^2 = 1 \end{aligned}$$

相関行列を

$$R = \begin{pmatrix} 1 & r_{xy} \\ r_{xy} & 1 \end{pmatrix}$$

とおくと，

$$\begin{pmatrix} s_X^2 & s_{XY} \\ s_{XY} & s_Y^2 \end{pmatrix} \equiv \begin{pmatrix} 1 & r_{xy} \\ r_{xy} & 1 \end{pmatrix} = R \tag{3.17}$$

となって，分散・共分散行列は相関行列 R に等しくなります。したがって (3.15) は

$$\begin{pmatrix} 1 & r_{xy} \\ r_{xy} & 1 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \tag{3.18}$$

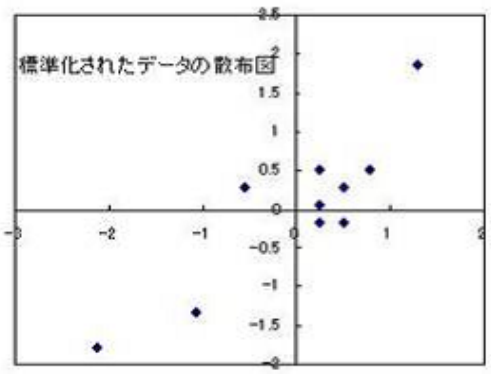
となります[18]。

以上のことを踏まえ，具体的な問題に取り組んでいくことにします。生徒 10 人の身長と体重の測定データを下表に示します。説明変量 x_1 として身長，x_2 として体重を取りますが，それぞれ単位が異なるのでデータを標準化しておきます。

No	身長(x1)	体重(x2)
1	172.0	63.0
2	173.0	64.0
3	172.0	61.0
4	171.0	62.0
5	171.0	64.0
6	166.0	56.0
7	162.0	54.0
8	168.0	63.0
9	171.0	61.0
10	175.0	70.0
平均	170.1	61.8
σ	3.7845	4.4171

標準化

	身長(x1)	体重(x2)
1	0.5021	0.2717
2	0.7663	0.4981
3	0.5021	-0.1811
4	0.2378	0.0453
5	0.2378	0.4981
6	-1.0834	-1.3131
7	-2.1403	-1.7658
8	-0.5549	0.2717
9	0.2378	-0.1811
10	1.2948	1.8564
分散	1.0	1.0
共分散	0.8854	



さて，第 1 主成分 z_1 の係数 a_1, a_2 を求めるのがここでの仕事になります。

$$z_1 = a_1 x_1 + a_2 x_2 \tag{3.19}$$

[17] セクション 1.4 参照。
[18] 説明変量の単位が異なっているときは相関行列による主成分分析を行うことがポイント。

標準化されたデータの相関係数（今の場合共分散と同値）は $r_{12} = 0.8854$ なので (3.32) より

$$\begin{pmatrix} 1.0 & 0.8854 \\ 0.8854 & 1.0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \tag{3.20}$$

の固有値問題を解くことになります。固有値と固有ベクトルの計算はパソコンで簡単にできて，結果は次の通りです[19]。

$$\begin{cases} \text{固有値：} & \lambda_1 = 1.8854 & \lambda_2 = 0.1146 \\ \text{固有ベクトル：} & a_1 = 0.7071, a_2 = 0.7071 & a_1 = 0.7071, a_2 = -0.7071 \end{cases} \tag{3.21}$$

固有値は主成分の分散に等しいので，主成分の分散を最大にする係数 a_1, a_2 は大きい方の固有値の固有ベクトルで，$a_1 = 0.7071$, $a_2 = 0.7071$ となって第 1 主成分は (3.29) より

$$z_1 = 0.7071x_1 + 0.7071x_2 \tag{3.22}$$

と求まります。ちなみに，小さい方の固有値は第 2 主成分となり，

$$z_2 = 0.7071x_1 - 0.7071x_2$$

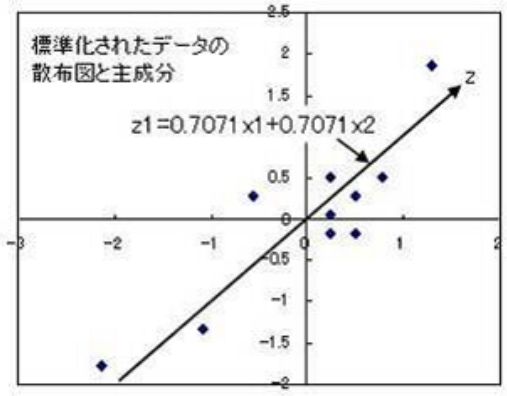
と表されます。尚，第 2 主成分についての詳細は後ほど触れることにします。

※蛇足：エクセルで標準化データの相関係数は容易に求りますが，以下に手計算での計算プロセスを示しておきます。主成分の分散を V と書くと，標準化されたデータの平均は 0 なので

$$\begin{aligned} V &= \frac{1}{n-1}\sum_{i=1}^{n}\{(p_i a_1 + q_i a_2) - (\bar{x}_1 a_1 + \bar{x_2} a_2)\}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(p_i a_1 + q_i a_2)^2 \\ &= \frac{1}{10-1}[(0.5021a_1 + 0.2717a_2)^2 + (0.7663a_1 + 0.4981a_2)^2 + \cdots + (1.2948a_1 + 1.8564a_2)^2 \\ &= a_1^2 + a_2^2 + 1.77071a_1a_2 = s_1^2 a_1^2 + s_2^2 a_2^2 + 2s_{12}a_1a_2 \\ \therefore \quad & s_1^2 = 1.0,\ s_2^2 = 1.0,\ s_{12} = 0.8854 \end{aligned}$$

※重要な蛇足：固有値 λ_1, λ_2 を足した値は説明変量の数に一致します。今の場合 $\lambda_1 + \lambda_2 = 1.8854 + 0.1146 = 2$ となって説明変量である身長と体重の個数に一致します。ただし，次ぎに述べる情報損失量を最小にする方法での固有値にはこのような性質はないので留意ください。　(蛇足終わり)

z_1 軸は $\tan\theta = a_2/a_1 = 1.0$ より $\theta = 45$ °の傾きを持つデータの中心を通る直線となりますから，これを図示すると下図のようになります。



[19] エクセルでの固有値計算プログラム→ http://www.qmss.jp/e-stat/excel/eigen.htm，フリーソフト Octave 入手先と使い方→ http://octave.futene.net/index.html を参照ください。

3.2.3 情報損失量を最小にする方法

次ぎに同じ資料で情報損失量を最小にする方法により第 1 主成分を求めます。点 (p_1, q_i) の情報損失量は (3.7) より

$$\frac{\mid a_2 p_i + a_1 q_i + a_0 \mid}{\sqrt{a_1^2 + a_2^2}} = \mid a_2 p_i + a_1 q_i + a_0 \mid \quad (\because a_1^2 + a_2^2 = 1) \tag{3.23}$$

となります。身長と体重のデータの情報損失量を次の表にまとめておきました。

No	身長 (x_1)	体重 (x_2)	情報損失量
1	0.5051	0.2717	$\mid 0.051a_2 - 0.2717a_1 + a_0 \mid$
2	0.7663	0.4981	$\mid 0.7663a_2 - 0.4981a_1 + a_0 \mid$
3	0.5021	-0.1811	$\mid 0.5021a_2 + 0.1811a_1 + a_0 \mid$
4	0.2378	0.0453	$\mid 0.2378a_2 - 0.0453a_1 + a_0 \mid$
5	0.2378	0.4981	$\mid 0.2378a_2 - 0.4981a_1 + a_0 \mid$
6	-1.0834	-1.3131	$\mid -1.0834a_2 + 1.3131a_1 + a_0 \mid$
7	-2.1403	-1.7658	$\mid -2.1403a_2 + 1.7658a_1 + a_0 \mid$
8	-0.5549	0.2717	$\mid -0.5549a_2 - 0.2717a_1 + a_0 \mid$
9	0.2378	-0.1811	$\mid 0.2378a_2 + 0.1811a_1 + a_0 \mid$
10	1.2948	1.8564	$\mid 1.2948a_2 - 1.8564a_1 + a_0 \mid$

情報損失量の平方の総和 $U(a_2, a_1, a_0)$ は

$$\begin{aligned} U(a_2, a_1, a_0) &= (0.5021a_2 - 0.2717a_1 + a_0)^2 + (0.7663a_2 - 0.4981a_1 + a_0)^2 \\ &\quad + \cdots + (1.2948a_2 - 1.8564a_1 + a_0)^2 \\ &= 9.0a_2^2 + 9.0a_1^2 - 15.9364a_1a_2 + 10a_0^2 \end{aligned} \tag{3.24}$$

となり，$a_2^2 + a_1^2 - 1 = 0$ という条件のもとで $U(a_2, a_1, a_0)$ の最小値を与える a_1, a_2 を求めることになります。次の新たな関数を導入します。

$$F = U(a_2, a_1, a_0) - \lambda(a_2^2 + a_1^2 - 1)$$

F の極値を与える a_2, a_1, a_0 は F をそれぞれの変数で偏微分したものが 0 となりますから

$$\frac{\partial F}{\partial a_2} = 0, \quad \frac{\partial F}{\partial a_1} = 0, \quad \frac{\partial F}{\partial a_0} = 0$$

この式と $a_2^2 + a_1^2 - 1 = 0$ を連立させて解けば係数が求まることになります。

$$\begin{cases} \dfrac{\partial F}{\partial a_2} = 18a_2 - 15.9364a_1 - 2\lambda a_2 = 0 \\ \dfrac{\partial F}{\partial a_1} = -15.9364a_2 + 18a_1 - 2\lambda a_1 = 0 \\ \dfrac{\partial F}{\partial a_0} = 20a_0 = 0 \end{cases} \tag{3.25}$$

(3.25) の上 2 式を整理すると

$$\begin{cases} (9 - \lambda)a_2 - 7.9682a_1 = 0 \\ -7.9682a_2 + (9 - \lambda)a_1 = 0 \end{cases} \Rightarrow \begin{pmatrix} 9 & -7.9682 \\ -7.9682 & 9 \end{pmatrix} \begin{pmatrix} a_2 \\ a_1 \end{pmatrix} = \lambda \begin{pmatrix} a_2 \\ a_1 \end{pmatrix} \tag{3.26}$$

となって，固有値問題に行き着きます。固有値と固有ベクトルは次のようになります。

$$\begin{cases} \text{固有値：} & \lambda_1 = 1.0318 & \lambda_2 = 16.9682 \\ \text{固有ベクトル：} & a_2 = 0.7071, a_1 = 0.7071 & a_2 = -0.7017, a_1 = 0.7071 \end{cases} \tag{3.27}$$

さて，分散を最大にする方法では“ 固有値は分散と同値 ”ということから，大きい固有値に属する固有ベクトルを求めればよかったわけですが，情報損失量を最大にする方法にはそのようなうまい求め方はないのでしょうか。以下にその辺りを調べてみると。。。情報損失量の平方の総和 (3.24) は

$$U = 9.0a_2^2 + 9.0a_1^2 - 15.9364a_2a_1 \tag{3.28}$$

また，(3.26) の左の第 1 式× a_2+ 第 2 式× a_1 を計算すると

$$9a_2^2 + 9a_1^2 - 15.9364a_2a_1 = \lambda(a_2^2 + a_1^2)$$

この式の左辺は (3.28) の U と等しいから

$$U = \lambda(a_2^2 + a_1^2) = \lambda \quad (\because a_2^2 + a_1^2 = 1) \tag{3.29}$$

となります。**情報損失量の平方の総和 $U(a_2, a_1, a_0)$ は固有値 λ に等しい** という重要な結果がでてきました[20]。そこで先ほどの固有値の選択方針は，最小の U に等しい最小の固有値を選択すればよいということになります。ということで (3.27) より求める第 1 主成分は

$$z_1 = 0.7071x_1 + 0.7071x_2 \tag{3.30}$$

となって，これは (3.22) と一致します。つまり いずれの方法も答えは一致 するということが確認できました... やれやれ。

3.2.4 主成分得点

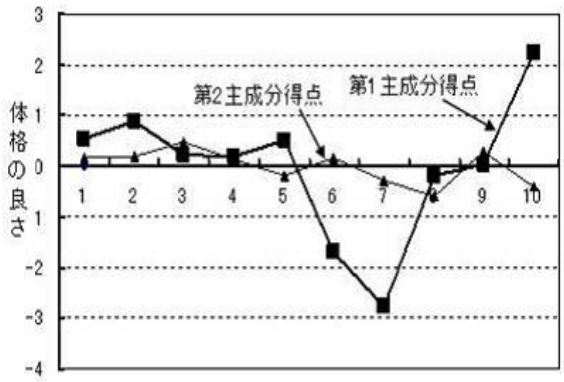
セクション 3.2.2 の相関行列を使った場合の主成分得点を求めていきます。第 1，2 主成分は (3.22)(3.23) より

$$\begin{cases} z_1 = 0.7071x_1 + 0.7071x_2 \\ z_2 = 0.7071x_1 - 0.7071x_2 \end{cases} \tag{3.31}$$

この式に標準化したデータを入れて第 1，2 の主成分得点を計算すると下表にようになります。

	身長(x1)	体重(x2)	第1主成分得点	第2主成分得点
1	0.5021	0.2717	0.5471	0.1629
2	0.7663	0.4981	0.8940	0.1897
3	0.5021	-0.1811	0.2269	0.4831
4	0.2378	0.0453	0.2002	0.1361
5	0.2378	0.4981	0.5203	-0.1840
6	-1.0834	-1.3131	-1.6945	0.1624
7	-2.1403	-1.7658	-2.7621	-0.2648
8	-0.5549	0.2717	-0.2003	-0.5845
9	0.2378	-0.1811	0.0401	0.2962
10	1.2948	1.8564	2.2282	-0.3971
		平均	0.0000	0.0000
		分散	1.8854	0.1146

固有値に一致する



主成分得点の分散は固有値に一致していることを確認ください。主成分得点を各生徒ごとにプロットしたものが右のグラフですが，元のデータ（身長と体重の散布図）を眺めるより各生徒の体格の特長を容易に把握できますね（2 次元データーを 1 次元に縮小！）。 第 1 主成分得点が最も高い生徒は N0.10 の生徒で，この生徒が最も体格のよい生徒と言えます。

[20] 分散を最大にするやり方では分散が固有値となり，情報損失量を最小にするやり方では情報損失量が固有値になります。この辺りは面白いですね

3.3　寄与率と累積寄与率

3.4　寄与率（主成分の情報収集能力）

主成分分析は主成分が与えられた資料の情報をどの程度説明しているのか，その辺りが重要になってきます。セクション 3.1 のところで「元の情報量の 2 乗和 = 新たな情報量の 2 乗和 + 情報損失量の 2 乗和」が成立することを述べましたが，主成分得点の絶対値を“ 新たな情報量 ”とすると，新たな情報量は元の情報量をどの程度説明しているのでしょうか。それを表す指標として

$$\text{主成分の寄与率} = \frac{\text{新たな情報量の 2 乗和}}{\text{元の情報量の 2 乗和}} = \frac{\text{元の情報量の 2 乗和} - \text{情報損失量の 2 乗和}}{\text{元の情報量の 2 乗和}}$$

を定義します。また，この式は (3.10) より次のようにも表わすことができます。

$$\text{主成分の寄与率} = \frac{\text{主成分の分散値}}{\text{各変量の分散の和}} \tag{3.32}$$

(3.32) を使って先ほどの資料の主成分の寄与率を求めると，各変量の分散は 1 で，第 1，2 主成分の分散値（固有値）はそれぞれ 1.8854，0.1146 であったので

$$\begin{cases} \text{第 1 主成分の寄与率} = \frac{1.8854}{1+1} = 0.9427 \\ \text{第 2 主成分の寄与率} = \frac{0.1146}{1+1} = 0.0573 \end{cases}$$

となり，第 1 主成分はデータの 94% を説明しており，第 2 主成分は残りの 6% を説明していることがわかります。これらの寄与率を足し合わせたものを**累積寄与率**と呼んでます。

3.5　累積寄与率

いまの場合は，身長と体重という 2 変量という単純なケースだったので第 1 主成分と第 2 主成分ですべてを説明することができましたが，一般に 3 変量以上の場合の寄与率についての定式化をやっておきます[21]。尚，以前書いた「重要な蛇足」のところも参照ください。
ポイントは

相関行列による主成分分析では，すべての固有値の合計は適用した変量の個数と一致する。

固有値を変量の個数で割った値を寄与率といい，第 i 主成分の寄与率は

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_n} = \frac{\lambda_i}{\sum_{k=1}^{n} \lambda_k} \tag{3.33}$$

で与えられる。

第 1 主成分から第 i 主成分までの累積寄与率は

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_n} = \frac{\sum_{k=1}^{i} \lambda_i}{\sum_{k=1}^{n} \lambda_k} \quad (i \leqq n) \tag{3.34}$$

となる，といったところでしょうか。寄与率は固有値が求まれば容易に計算することができるというわけですね。

主成分分析の計算はパソコン（エクセル）を使えば簡単に結果を得ることができますが，結果の解釈は解析者に任されるので，いかにうまい解釈をするかが重要なポイントとなります。

[21] 証明はここではやらないので証明が欲しい方は多変量解析の適当なテキストに当たってみてください。例えば永田靖・棟近雅彦「多変量解析法入門」（サイエンス社）等。

3.6　主成分分析の例題

	動力性能(x1)	居住性(x2)	デザイン(x3)
A	60	58	25
B	35	40	75
C	74	68	50
D	30	40	60
E	80	70	50
F	90	95	80
G	50	50	45
平均	60	60	55
分散	520	382	350
s_12	427		
s_23	74		
s_13	5		

それでは主成分分析の項を終わるにあたって最後に例題[22]をやってみます。右の表は同じ値段の 7 台の車についての 100 人のアンケート結果で，中の数字は「よい」と評価した人の数です。この資料を主成分分析して各車の特性はどのようなものかを見てみましょう。主成分分析には「分散・共分散行列」を利用する場合と「相関行列」を利用する場合がありますが，ここでは「分散・共分散行列」を使った分析をやることにします[23]。

主成分が説明変量 x_1(動力性能) , x_2(居住性) , x_3(デザイン) の一次式

$$z = a_1x_1 + a_2x_2 + a_3x_3$$

で表されるとします。変量 x_1, x_2, x_3 の分散・共分散はエクセルで容易に計算でき，分散・共分散行列は

$$S = \begin{pmatrix} 520 & 427 & 5 \\ 427 & 382 & 74 \\ 5 & 74 & 350 \end{pmatrix} \tag{3.35}$$

となり，この行列の固有値とそれに属する固有ベクトルとして

$$\lambda_1 = 889, (0.75, 0.65, 0.1), \quad \lambda_2 = 353, (-0.17, 0.06, 0.98), \quad \lambda_3 = 10, (-0.63, 0.75, -0.16)$$

が得られます。最大となる固有値 889 が主成分の分散になり，第 1 主成分は

$$z_1 = 0.75x_1 + 0.65x_2 + 10x_3 \tag{3.36}$$

と表され，同様にして第 2 主成分は

$$z_2 = -0.17x_1 + 0.06x_2 + 0.98x_3 \tag{3.37}$$

となります。

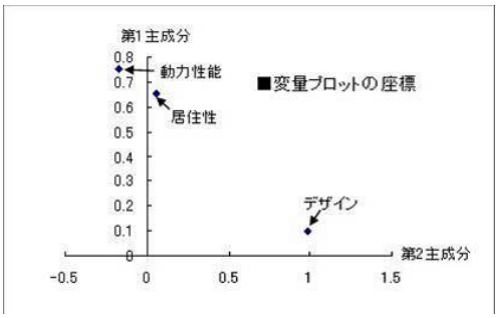
■変量プロットの座標

そこで縦軸に第 1 主成分，横軸に第 2 主成分をとり，変量をこの座標上にプロットすると右図のようになります。この図から

第 1 主成分は各変量がプラスに評価されており，車の「総合評価」の視点を提供している

第 2 主成分は動力性能とデザインが逆でデザインが大きく評価されており，車の「ファッション性」を評価する視点を提供している

ことが分かります。



■主成分得点

次ぎに，主成分得点を求めっていきます。

分散・共分散行列を利用する主成分分析の場合，i 番目のサンプルの主成分得点は (x_1^i, x_2^i, x_3^i) は i 番目のサンプルの変量 x_1, x_2, x_3 の値として

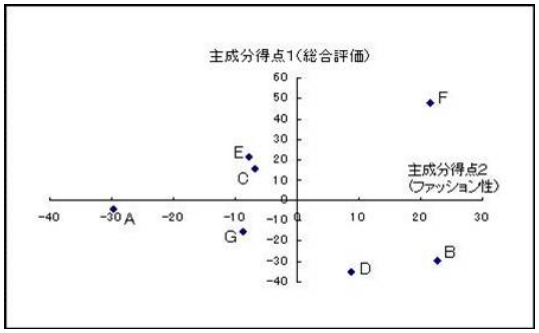
$$a_1(x_1^i - \bar{x}_1) + a_2(x_2^i - \bar{x}_2) + a_3(x_3^i - \bar{x}_3) \tag{3.38}$$

	主成分得点1	主成分得点2
A	-4.3	-29.6
B	-29.7	22.6
C	15.2	-6.8
D	-35.0	8.8
E	21.0	-7.7
F	47.8	21.5
G	-15.0	-8.7

[22] 涌井良幸，涌井貞美「図解でわかる多変量解析」より引用。
[23] 相関行列の利用は各自フォローされたし。

で表され，これを計算した結果が右の表となります。第 1
主成分は「総合評価」，第 2 主成分は「ファッション性」を示していましたから，例えば車 B は「総合評価は低いがファッション性は高い」ということがこの表からすぐに判断できます。

主成分得点を図示したものが右図です。このセクションの一番最初のデータと見比べてください。一番最初のデータを眺めても各車の特性はなかなか掴みにくいですが，主成分得点を図示した図からは車 A〜F の特性が一目瞭然に分かりますね。



$= \mathscr{MEMO} =$

■ラグランジュの未定係数法

束縛（拘束）条件 $g(x_1, x_2, \cdots, x_n) = 0$ のもとで，
関数 $z = f(x_1, x_2, \cdots, x_n)$ が極値を持つとき，その
極値点を $(x_1, x_2, \cdots, x_n)$ とすると，極値点は新たな関数 $G(x_1, x_2, \cdots, x_n, \lambda)$ を

$$G(x_1, x_2, \cdots, x_n, \lambda) = f(x_1, x_2, \cdots, x_n) - \lambda g(x_1, x_2, \cdots, x_n)$$

とおいて，G を $x_1, x_2, \cdots, x_n$ で偏微分したものが 0 になる値として求めることができます。

$$\begin{cases} \dfrac{\partial G}{\partial x_1} = \dfrac{\partial f}{\partial x_1} - \lambda \dfrac{\partial g}{\partial x_1} = 0 \\ \dfrac{\partial G}{\partial x_2} = \dfrac{\partial f}{\partial x_2} - \lambda \dfrac{\partial g}{\partial x_2} = 0 \\ \qquad \vdots \\ \dfrac{\partial G}{\partial x_n} = \dfrac{\partial f}{\partial x_n} - \lambda \dfrac{\partial g}{\partial x_n} = 0 \end{cases} \tag{3.39}$$

求める極値点 $(x_1, x_2, \cdots, x_n y)$ は束縛条件 $g(x_1, x_2, \cdots, x_n) = 0$ と（3.39）から求まります。

＜ラグランジュの未定乗数法の詳しい説明＞　（とくに詳しい説明不要の方は飛ばしても OK）
簡単のために独立変数が x, y, z の 3 つである場合を考えると，3 変数の関数 $f(x, y, z)$ の極値問題はセクション 1.5「単回帰分析」のところでやったように

$$\frac{\partial f(x, y, z)}{\partial x} = 0, \quad \frac{\partial f(x, y, z)}{\partial y} = 0, \quad \frac{\partial f(x, y, z)}{\partial z} = 0 \tag{3.40}$$

の連立方程式を解けばよかったわけです。ところが，x, y, z がそれぞれ独立ではなく，束縛条件 $g(x, y, z) = 0$ がある場合には上のように簡単にはいきません。というのは，(3.40) の条件は，関数 $f(x, y, z)$ の全微分

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \frac{\partial f}{\partial z} dz \tag{3.41}$$

が極値点の周りに x, y, z が独立して微少量 dx, dy, dz 変化しても 0 という条件[24]からきていますが，束縛条件があると微少量 dx, dy, dz を勝手に動かすことはできません。極値点で束縛条件が成り立っているとすると

$$g(x + dx, y + dy, z + dz) = g(x, y, z)$$

を満たす変化しか許されないわけですね。では許される変化はどのようなものかというと，上の式は右辺の項を左辺に移項すると関数 $g(x, y, z)$ の全微分 dg に相当しますから

$$dg = g(x + dx, y + dy, z + dz) - g(x, y, z) = \frac{\partial g}{\partial x} dx + \frac{\partial g}{\partial y} dy + \frac{\partial g}{\partial z} dz = 0$$

これを dz について解くと

$$dz = -\frac{\dfrac{\partial g}{\partial x}dx + \dfrac{\partial g}{\partial y}dy}{\dfrac{\partial g}{\partial z}} = -\left\{ \left(\frac{\dfrac{\partial g}{\partial x}}{\dfrac{\partial g}{\partial z}} \right) dx + \left(\frac{\dfrac{\partial g}{\partial y}}{\dfrac{\partial g}{\partial z}} \right) dy \right\}$$

が得られます。したがって z の微小変化 dz は x, y の微小変化 dx, dy を与えると決まるという構造になります。これを (3.41) に入れると

$$\begin{aligned} df &= \left(\frac{\partial f}{\partial x} - \frac{\dfrac{\partial g}{\partial x}}{\dfrac{\partial g}{\partial z}} \frac{\partial f}{\partial z} \right) dx + \left(\frac{\partial f}{\partial y} - \frac{\dfrac{\partial g}{\partial y}}{\dfrac{\partial g}{\partial z}} \frac{\partial f}{\partial z} \right) dy \\ &= \left(\frac{\partial f}{\partial x} - \lambda \frac{\partial g}{\partial x} \right) dx + \left(\frac{\partial f}{\partial y} - \lambda \frac{\partial g}{\partial y} \right) dy \\ &= 0\,, \qquad \left(\text{ただし}, \lambda = \frac{\partial f}{\partial z} \Big/ \frac{\partial g}{\partial z} \Longrightarrow \frac{\partial f}{\partial z} - \lambda \frac{\partial g}{\partial z} = 0 \right) \end{aligned}$$

となり，dx と dy の 2 個だけであれば自由に変えることができるので，(3.42) の恒等式が成り立つためには，右辺の各係数は 0 でなければなりません。つまり，

$$\frac{\partial f}{\partial x} - \lambda \frac{\partial g}{\partial x} = 0, \quad \frac{\partial f}{\partial y} - \lambda \frac{\partial g}{\partial y} = 0$$

ということで，結局，束縛条件付の極値問題は

$$G = f - \lambda g \tag{3.42}$$

とおいて，x, y, z が独立に変化しうるものと考えて $h(x, y, z)$ の極値問題を解くことと同等である，ということになります。未知数は x, y, z と λ の 4 つで，これらは次の 4 つの方程式を連立させて解くことで求めることができます。なお，λ は**ラグランジュの未定乗数**と呼ばれています。

$$\begin{cases} g(x, y, z) = 0 \\ \dfrac{\partial f}{\partial x} - \lambda \dfrac{\partial g}{\partial x} = 0 \\ \dfrac{\partial f}{\partial y} - \lambda \dfrac{\partial g}{\partial y} = 0 \\ \dfrac{\partial f}{\partial z} - \lambda \dfrac{\partial g}{\partial z} = 0 \end{cases} \tag{3.43}$$

4 因子分析

4.1 因子分析の概要

世の中の現象はいろいろな“ 要因 ”が絡まりあって起こっています。いま，あるクラスで数学，国語，英語の試験が実施され，A，B，C，D の 4 人が下表のような得点を得たものとしましょう。

No.	数学	国語	英語
A	30	50	45
B	70	60	60
C	45	55	55
D	40	35	50

さて,「各学生は“ 理系能力因子 ”と“ 文系能力因子 ”を持っており，この 2 つの因子の大小が数学，国語，英語の試験の点数を決定する」という仮定を設けます。尤も，数学の得点が高いから理系能力因子の寄与が大きいとは言えても，問題の読解には文系能力因子も影響していると考えられます。同様に国語や英語の試験においても，問題の大意を掴むようなときには頭の中で収束性のある文を構成しなければならず，理系能力も要求されるでしょう。さらに全般的には各学生の勉強時間や今まで蓄積してきた知識なども試験の得点に反映されてくるでしょう。ここでは，勉強時間の長短や今までの知識などの要因は“ **誤差** ”として取り扱うことにします。

各学生の理系能力因子と文系能力因子が次のような数値（これを**因子得点**と呼んでいます）で定まっているものとしましょう[25]。

＜共通因子＞

No.	数学	国語	英語		理系能力因子	文系能力因子
A	30	50	45		20	40
B	70	60	60	$\Longrightarrow$	60	40
C	45	55	55		30	50
D	40	35	50		50	20

（※：共通因子内の数値は**因子得点**）

ここで注意しなければならない点は，理系能力因子と文系能力因子は互いに相関のない全く異なった能力因子であるという点です。このように仮定した因子を**直交因子**と呼んでいます。また，これらの因子は，その大きさは別として，各学生が共通して持っている因子という意味で**共通因子**とも呼ばれます。

さて，各学科の得点と理系・文系能力因子を関係付けていきましょう。因子を得点に反映させるために次の係数を導入します。

“ 数学 ”という変数が理系能力因子に作用する大きさ　：　a_{m1}
“ 数学 ”という変数が文系能力因子に作用する大きさ　：　a_{m2}
“ 国語 ”という変数が理系能力因子に作用する大きさ　：　$a_{\ell 1}$
“ 国学 ”という変数が文系能力因子に作用する大きさ　：　$a_{\ell 2}$
“ 英語 ”という変数が理系能力因子に作用する大きさ　：　a_{e1}
“ 英語 ”という変数が文系能力因子に作用する大きさ　：　a_{e2}

これらの係数を**因子負荷量**と呼んでいます。因子負荷量のイメージは少し捉えにくいかも知れないので，そのの働きを以下に補足しておきます。

$\left\{\begin{array}{ll} a_{m1} & : \text{数学という変数が理系能力因子に作用して理系能力因子から数学関係する部分をとってくる。} \\ a_{m2} & : \quad 〃 \quad \text{文系能力因子} \quad 〃 \quad \text{文系能力因子から数学関係する部分} \quad 〃 \\ a_{\ell 1} & : \text{国語という変数が理系能力因子に作用して理系能力因子から国語関係する部分をとってくる。} \\ a_{\ell 2} & : \quad 〃 \quad \text{文系能力因子} \quad 〃 \quad \text{文系能力因子から国語関係する部分} \quad 〃 \\ a_{e1} & : \text{英語という変数が理系能力因子に作用して理系能力因子から英語関係する部分をとってくる。} \\ a_{e2} & : \quad 〃 \quad \text{文系能力因子} \quad 〃 \quad \text{文系能力因子から英語関係する部分} \quad 〃 \end{array}\right.$

[25] 因子得点は直接観測することができない潜在変数で，実際には因子分析を行って求まります。

そこで，例えば学生 A の数学・国語・英語の実際の得点を文系・理系能力因子（具体的には因子得点）と因子負荷量でもって表すと，理論上の得点に誤差 ε を加えて

$$\begin{cases} \text{学生 } A \text{ の数学の得点} & : \quad 30 = a_{m1} \times 20 + a_{m2} \times 40 + \varepsilon_{Am} \\ 〃 \quad \text{国語の得点} & : \quad 50 = a_{\ell 1} \times 20 + a_{\ell 2} \times 40 + \varepsilon_{A\ell} \\ 〃 \quad \text{英語の得点} & : \quad 45 = a_{e1} \times 20 + a_{e2} \times 40 + \varepsilon_{Ae} \end{cases} \tag{4.1}$$

というように表すことができます[26]。いま因子負荷量を仮に次のように設定してみます。

a_{m1}	a_{m2}	$a_{\ell 1}$	$a_{\ell 2}$	a_{e1}	a_{e2}
0.9	0.3	0.4	0.85	0.6	0.7

そうすると，各学生の数学，国語，英語の理論上の得点は，(4.1) から分かるように

■数学の理論上の得点

	理系能力因子	文系能力因子	理論上の得点	実際の得点	誤差
A	20	40	0.9×20+0.3×40=30	30	0
B	60	40	0.9×60+0.3×40=66	70	4
C	30	50	0.9×30+0.3×50=42	45	3
D	50	20	0.9×50+0.3×20=46	40	-6

■国語の理論上の得点

	理系能力因子	文系能力因子	理論上の得点	実際の得点	誤差
A	20	40	0.4×20+0.85×40=42	50	8
B	60	40	0.4×60+0.85×40=58	60	2
C	30	50	0.4×30+0.85×50=54.5	55	0.5
D	50	20	0.4×50+0.85×20=37	35	-2

■英語の理論上の得点

	理系能力因子	文系能力因子	理論上の得点	実際の得点	誤差
A	20	40	0.6×20+0.7×40=40	45	5
B	60	40	0.6×60+0.7×40=64	60	-4
C	30	50	0.6×30+0.7×50=53	55	2
D	50	20	0.6×50+0.7×20=44	50	6

以上が因子分析の概要です。整理すると次のようになります。
因子分析とは単純な要因で複雑なものを説明しようとする統計的な手法のことで，ある現象（データ）を捉えたとき，その現象が仮に一つの要因（因子）で起こっているとすると，その現象は次の線形モデルで示すことができると考えます。

$$\text{現象（データ）} = \text{因子負荷量} \times (\text{因子得点}) + \text{誤差}$$

考えられる因子が 2 個，あるいは一般化して n 個あるとするなら，上の式は

$$\begin{cases} \text{現象（データ）} = \text{因子負荷量 } 1 \times (\text{因子得点 } 1) + \text{因子負荷量 } 2 \times (\text{因子得点 } 2) + \text{誤差} \\ \text{現象（データ）} = \text{因子負荷量 } 1 \times (\text{因子得点 } 1) + \text{因子負荷量 } 2 \times (\text{因子得点 } 2) + \cdots \\ \qquad\qquad + \text{因子負荷量 } n \times (\text{因子得点 } n) + \text{誤差} \end{cases}$$

[26] 先ほどの理系・文系能力因子は各学生個々人の属性値であったが，因子負荷量は学生個々人の属性値ではなく，それら因子に負荷する量であるということに留意されたし。

と表せます。

具体的に，n 人の生徒の試験の成績データー（右表）を使って今までの話を数学的に整理していきましょう。国語，英語，数学，理科をそれぞれ変量 z_1, z_2, z_3, z_4 とし，表の枠内の値は各生徒の学科試験の得点とします。尚，変量は**標準化**されているとします[27]。そうするとNo.k番目の生徒の試験の成績は次式で表すことができます。

生徒No	国語(z1)	英語(z2)	数学(z3)	理科(z4)
1	z11	z12	z13	z14
2	z21	z22	z23	z24
3	z31	z32	z33	z34
:	:	:	:	:
k	zk1	zk2	zk3	zk4
:	:	:	:	:
n	zn1	zn2	zn3	zn4

$$\left\{\begin{array}{l} \text{国語の得点：} z_{k1} = a_{11} \times (\text{文系能力因子得点}) + a_{12} \times (\text{理系能力因子得点}) + \text{誤差} \\ \text{英語　〃　：} z_{k2} = a_{21} \times (\text{文系能力因子得点}) + a_{22} \times (\text{理系能力因子得点}) + \text{誤差} \\ \text{数学　〃　：} z_{k3} = a_{31} \times (\text{文系能力因子得点}) + a_{32} \times (\text{理系能力因子得点}) + \text{誤差} \\ \text{理科　〃　：} z_{k4} = a_{41} \times (\text{文系能力因子得点}) + a_{42} \times (\text{理系能力因子得点}) + \text{誤差} \end{array}\right. \tag{4.2}$$

右辺は試験の得点で実測値ですが，左辺の量はすべて観測できない量であるという点に留意してください。

生徒の共通因子と因子得点を右表にまとめました。f_{k1}, f_{k2} 等は**因子得点**ですね。“ **誤差** ”は共通因子だけでは説明できない独自の要素のことで，この部分は特に**独自因子**とか**独自部分**と呼ばれています。

整理すると全生徒の各教科の成績は次の $4n$ 個の式で表されることになります。これをとりあえず“ 因子分析の数学モデル ”と呼ぶことにします。

生徒No	文系的能力	理系的能力
1	f11	f12
2	f21	f22
3	f31	f32
:	:	:
k	fk1	fk2
:	:	:
n	fn1	fn2

$$i\text{番目の生徒の成績}\left\{\begin{array}{lll} \text{国語の得点：} & z_{i1} = a_{11}f_{i1} + a_{12}f_{i2} + \epsilon_{i1} & \\ \text{英語　〃　：} & z_{i2} = a_{21}f_{i1} + a_{22}f_{i2} + \epsilon_{i2} & \\ \text{数学　〃　：} & z_{i3} = a_{31}f_{i1} + a_{32}f_{i2} + \epsilon_{i3} & \\ \text{理科　〃　：} & z_{i4} = a_{41}f_{i1} + a_{42}f_{i2} + \epsilon_{i4} & (\text{生徒の数：} i = 1, 2, \cdots, n) \end{array}\right. \tag{4.3}$$

全体を見通せるように，上の式を行列形式で書きなおすと

$$\begin{pmatrix} z_{11} & z_{12} & z_{13} & z_{14} \\ z_{21} & z_{22} & z_{23} & z_{24} \\ \vdots & \vdots & \vdots & \vdots \\ z_{n1} & z_{n2} & z_{n3} & z_{n4} \end{pmatrix} = \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \\ \vdots & \vdots \\ f_{n1} & f_{n2} \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} & a_{31} & a_{41} \\ a_{12} & a_{22} & a_{32} & a_{42} \end{pmatrix} + \begin{pmatrix} e_{11} & e_{12} & e_{13} & e_{14} \\ e_{21} & e_{22} & e_{23} & e_{24} \\ \vdots & \vdots & \vdots & \vdots \\ e_{n1} & e_{n2} & e_{n3} & e_{n4} \end{pmatrix} \tag{4.4}$$

となります。そこで

$$\boldsymbol{Z} = \begin{pmatrix} z_{11} & z_{12} & z_{13} & z_{14} \\ z_{21} & z_{22} & z_{23} & z_{24} \\ \vdots & \vdots & \vdots & \vdots \\ z_{n1} & z_{n2} & z_{n3} & z_{n4} \end{pmatrix}, \quad F = \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \\ \vdots & \vdots \\ f_{n1} & f_{n2} \end{pmatrix}$$

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \end{pmatrix}, \qquad \boldsymbol{E} = \begin{pmatrix} e_{11} & e_{12} & e_{13} & e_{14} \\ e_{21} & e_{22} & e_{23} & e_{24} \\ \vdots & \vdots & \vdots & \vdots \\ e_{n1} & e_{n2} & e_{n3} & e_{n4} \end{pmatrix}$$

[27] 標準化しておくと単位のそろっていないデータも取り扱え，また数学的な取り扱いの見通しが良くなります。

という行列を定義すると，(4.4) は

$$\boldsymbol{Z} = \boldsymbol{F}\boldsymbol{A}^t + \boldsymbol{E} \tag{4.5}$$

と大変コンパクトな式になります。行列$\boldsymbol{A}$を**因子負荷行列**と呼んでいます。A^t は行列 A の転置行列[28]です。

4.2 因子分析の計算

因子分析の計算は，与えられたデータから因子の存在を想定して因子負荷量を求め，個々のサンプルの因子得点を算出するということになります。以下では因子分析計算の概要プロセスを解説します（←実際の計算は手計算では非常に面倒なため，適当な統計解析ソフトが使われています）。

4.2.1 変量の分散式より

それでは，前出 (4.3) の因子分析の数学モデルの構造を調べていきましょう。

$$\begin{cases} \text{国語の得点：} & z_{i1} = a_{11}f_{i1} + a_{12}f_{i2} + \epsilon_{i1} \\ \text{英語　〃　：} & z_{i2} = a_{21}f_{i1} + a_{22}f_{i2} + \epsilon_{i2} \\ \text{数学　〃　：} & z_{i3} = a_{31}f_{i1} + a_{32}f_{i2} + \epsilon_{i3} \\ \text{理科　〃　：} & z_{i4} = a_{41}f_{i1} + a_{42}f_{i2} + \epsilon_{i4} \qquad (\text{生徒の数：} i = 1, 2, \cdots, n) \end{cases} \tag{4.6}$$

この方程式は変数の数が多くとっつきにくいので，変数の数を減らすことができないかトライしてみます。国語，英語，数学，理科の各変量は標準化されているので，平均は 0，分散は 1 となりますね[29]。ここを突破口にしてやってみます。ここでは国語を取り上げて計算を進めます（英語，数学，理科も同様に計算できる）。

(1) 変量の分散は１

$$\text{「国語」の分散} = \frac{1}{n-1}(z_{11}^2 + z_{21}^2 + \cdots + z_{n1}^2) = 1 \rightarrow z_{11}^2 + z_{21}^2 + \cdots + z_{n1}^2 = n - 1 \tag{4.7}$$

これを (4.6) の第一式に入れると

$$\begin{aligned} &(a_{11}f_{11} + a_{12}f_{12} + e_{11})^2 + (a_{11}f_{21} + a_{12}f_{22} + e_{21})^2 + \cdots + (a_{11}f_{n1} + a_{12}f_{n2} + e_{n1})^2 = n - 1 \\ \rightarrow\ & a_{11}^2(f_{11}^2 + f_{21}^2 + \cdots + f_{n1}^2) + a_{12}^2(f_{12}^2 + f_{22}^2 + \cdots + f_{n2}^2) \\ & \quad +2a_{11}a_{12}(f_{11}f_{12} + f_{21}f_{22} + \cdots + f_{n1}f_{n2}) + (e_{11}^2 + e_{21}^2 + \cdots + e_{n1}^2) \\ & \quad +2a_{11}(f_{11}e_{11} + f_{21}e_{21} + \cdots + f_{n1}e_{n1}) + 2a_{12}(f_{12}e_{11} + f_{22}e_{21} + \cdots + f_{n2}e_{n1}) = n - 1 \end{aligned} \tag{4.8}$$

が得られます。(4.8) を次の (1)〜(4) を通して簡略化していきます。

(2) 因子得点の標準化

(4.8) の第 1,2 項に注目しましょう。

$$\begin{cases} f_{11}^2 + f_{21}^2 + \cdots + f_{n1}^2 \\ f_{12}^2 + f_{22}^2 + \cdots + f_{n2}^2 \end{cases} \tag{4.9}$$

[28] 転置行列とは行列の行と列を入れ替えた行列のことです。
[29] セクション 1.4「標準化」を参照。

の 2 式は因子得点の 2 乗和を表しています。各生徒の因子得点の測り方については特に何も既定していませんでしたので，因子得点の標準化を行って数学的取り扱いの見通しをよくします。因子得点を標準化することで

$$\begin{cases} \text{平均：} & f_{11}+f_{21}+\cdots+f_{n1}=0, \quad f_{12}+f_{22}+\cdots+f_{n2}=0 \\ \text{分散：} & \dfrac{1}{n-1}(f_{11}^2+f_{21}^2+\cdots+f_{n1}^2)=1, \quad \dfrac{1}{n-1}(f_{12}^2+f_{22}^2+\cdots+f_{n2}^2)=1 \end{cases} \tag{4.10}$$

という関係式が得られます。

(3) 共通因子の独立性

因子得点は標準化されているので，(4.8) の 2 行目第 1 項の式 $f_{11}f_{12}+f_{21}f_{22}+\cdots+f_{n1}f_{n2}$ は因子得点の共分散（または相関係数）に比例したものとなります。ところで，因子得点の間には相関はない[30]としたので共分散は 0 となります。つまり，

$$\begin{cases} \text{共分散：} \dfrac{1}{n-1}(f_{11}f_{12}+f_{21}f_{22}+\cdots+f_{n1}f_{n2})=0 \\ \therefore \quad f_{11}f_{12}+f_{21}f_{22}+\cdots+f_{n1}f_{n2}=0 \end{cases} \tag{4.11}$$

大分見通しが良くなってきました。次ぎに

(4) 独自因子の平均は 0

変量，因子得点の平均はそれぞれ 0 ですから，独自因子の平均も 0 になりますね。独自因子の分散をそれぞれ $d_1^2, d_2^2, d_3^2, d_4^2$ と置くと，

$$\begin{cases} d_1^2=\dfrac{1}{n-1}(e_{11}^2+e_{21}^2+\cdots+e_{n1}^2) \\ d_2^2=\dfrac{1}{n-1}(e_{12}^2+e_{22}^2+\cdots+e_{n2}^2) \\ d_3^2=\dfrac{1}{n-1}(e_{13}^2+e_{23}^2+\cdots+e_{n3}^2) \\ d_4^2=\dfrac{1}{n-1}(e_{14}^2+e_{24}^2+\cdots+e_{n4}^2) \end{cases} \longrightarrow \begin{cases} e_{11}^2+e_{21}^2+\cdots+e_{n1}^2=(n-1)d_1^2 \\ e_{12}^2+e_{22}^2+\cdots+e_{n2}^2=(n-1)d_2^2 \\ e_{13}^2+e_{23}^2+\cdots+e_{n3}^2=(n-1)d_3^2 \\ e_{14}^2+e_{24}^2+\cdots+e_{n4}^2=(n-1)d_4^2 \end{cases} \tag{4.12}$$

が得られます。最後に，

(4) 共通因子と独自因子は無相関

(4.8) の最後の項 $f_{11}e_{11}+f_{21}e_{21}+\cdots+f_{n1}e_{n1}, f_{12}e_{11}+f_{22}e_{21}+\cdots+f_{n2}e_{n1}$ は共通因子と独自因子の共分散（または相関係数）に比例しますが，共通因子と独自因子は無相関と仮定したので，その共分散は 0 となって，結局次式が成立します。

$$\begin{cases} f_{11}e_{11}+f_{21}e_{21}+\cdots+f_{n1}e_{n1}=0 \\ f_{12}e_{11}+f_{22}e_{21}+\cdots+f_{n2}e_{n1}=0 \end{cases} \tag{4.13}$$

長かった準備も以上で終わります。(4.10)〜(4.13) を (4.8) に入れて整理すると

$$(n-1)a_{11}^2+(n-1)a_{12}^2+(n-1)d_1^2=n-1 \rightarrow a_{11}^2+a_{12}^2+d_1^2=1$$

と大変シンプルな式に還元できました。他の変量も同様に計算することができて，結果をまとめて書くと

$$a_{i1}^2+a_{i2}^2+d_i^2=1 \rightarrow \begin{cases} a_{11}^2+a_{12}^2+d_1^2=1 \\ a_{21}^2+a_{22}^2+d_2^2=1 \\ a_{31}^2+a_{32}^2+d_3^2=1 \\ a_{41}^2+a_{42}^2+d_4^2=1 \end{cases} \tag{4.14}$$

となります。

[30] 文系能力と理系能力の間には相関はないと仮定します。

4.2.2 相関行列

次ぎに，変量間の共分散を調べてみましょう。変量は標準化されているので，平均は 0，分散は 1 ですね。また，共分散は相関係数と一致します。例えば国語と英語の相関係数を r_{12} とすると

$$r_{12} = \frac{1}{n-1}(z_{11}z_{12} + z_{21}z_{22} + \cdots + z_{n1}z_{n2}) \tag{4.15}$$

右辺の括弧内の式に (4.6) の前半 2 式を入れて整理し，今まで得られた結果を適用すると

$$\begin{aligned}
&z_{11}z_{12} + z_{21}z_{22} + \cdots + z_{n1}z_{n2} \\
&\quad = (a_{11}f_{11} + a_{12}f_{12} + e_{11})(a_{21}f_{11} + a_{22}f_{12} + e_{12}) \\
&\qquad +(a_{11}f_{21} + a_{12}f_{22} + e_{21})(a_{21}f_{21} + a_{22}f_{22} + e_{22}) + \cdots \\
&\qquad +(a_{11}f_{n1} + a_{12}f_{n2} + e_{n1})(a_{21}f_{n1} + a_{22}f_{n2} + e_{n2}) \\
&\quad = a_{11}a_{21}(f_{11}^2 + f_{21}^2 + \cdots + f_{n1}^2) + a_{12}a_{22}(f_{12}^2 + f_{22}^2 + \cdots + f_{n2}^2) \\
&\qquad +(e_{11}e_{12} + e_{21}e_{22} + \cdots + e_{n1}e_{n2}) \\
&\qquad +(a_{11}a_{22} + a_{12}a_{21})(f_{11}f_{12} + f_{11}f_{12} + \cdots + f_{11}f_{12}) \\
&\qquad +a_{11}(f_{11}e_{11} + f_{21}e_{22} + \cdots + f_{n1}e_{n2}) + a_{12}(f_{12}e_{12} + f_{22}e_{22} + \cdots + f_{n2}e_{n2}) \\
&\qquad +a_{21}(f_{11}e_{11} + f_{21}e_{21} + \cdots + f_{n1}e_{n1}) + a_{22}(f_{12}e_{11} + f_{22}e_{21} + \cdots + f_{n2}e_{n1}) \\
&\quad = a_{11}a_{21}(n-1) + a_{12}a_{22}(n-1) + (e_{11}e_{12} + e_{21}e_{22} + \cdots + e_{n1}e_{n2})
\end{aligned}$$

となり，相関係数は

$$r_{12} = a_{11}a_{21} + a_{12}a_{22} + \frac{1}{n-1}(e_{11}e_{12} + e_{21}e_{22} + \cdots + e_{n1}e_{n2})$$

と表すことができます。ところで上の式の右辺第 2 項は，変量「国語」と「英語」の独自部分の共分散を表しますが，これは互いに無相関と考えられる（←仮定）ので 0 と置けます。これから

$$r_{12} = a_{11}a_{21} + a_{12}a_{22}$$

が得られます。まったく同様にして，各変量間の相関係数 $r_{ij}(i < j)$ をまとめて書くと

$$r_{ij} = a_{i1}a_{j1} + a_{i2}a_{j2} \rightarrow \left\{ \begin{array}{ll} r_{12} = a_{11}a_{21} + a_{12}a_{22}, & r_{13} = a_{11}a_{31} + a_{12}a_{32} \\ r_{14} = a_{11}a_{41} + a_{12}a_{42}, & r_{23} = a_{21}a_{31} + a_{22}a_{32} \\ r_{24} = a_{21}a_{41} + a_{22}a_{42}, & r_{34} = a_{31}a_{41} + a_{32}a_{42} \end{array} \right. \tag{4.16}$$

と表すことができます。ただし $r_{ij} = r_{ji}$。これらの結果を行列形式でまとめると

$$\boldsymbol{R} = \begin{pmatrix} 1 & r_{12} & r_{13} & r_{14} \\ r_{21} & 1 & r_{23} & r_{24} \\ r_{31} & r_{32} & 1 & r_{34} \\ r_{41} & r_{42} & r_{43} & 1 \end{pmatrix}, \quad \boldsymbol{Z} = \begin{pmatrix} z_{11} & z_{12} & z_{13} & z_{14} \\ z_{21} & z_{22} & z_{23} & z_{24} \\ \vdots & \vdots & \vdots & \vdots \\ z_{n1} & z_{n2} & z_{n3} & z_{n4} \end{pmatrix}$$

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \end{pmatrix}, \qquad \boldsymbol{D} = \begin{pmatrix} d_1^2 & 0 & 0 & 0 \\ 0 & d_2^2 & 0 & 0 \\ 0 & 0 & d_3^2 & 0 \\ 0 & 0 & 0 & d_4^2 \end{pmatrix}$$

として，(4.15) を導出する相関行列 $\boldsymbol{R}$ は

$$\boldsymbol{R} = \frac{1}{n-1}\boldsymbol{Z}^t\boldsymbol{Z} \tag{4.17}$$

と表せます。また，(4.16) より，相関行列 $\boldsymbol{R}$は $\boldsymbol{A}$, $\boldsymbol{D}$ を使って次のように表すことができます。

$$\begin{pmatrix} 1 & r_{12} & r_{13} & r_{14} \\ r_{21} & 1 & r_{23} & r_{24} \\ r_{31} & r_{32} & 1 & r_{34} \\ r_{41} & r_{42} & r_{43} & 1 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} & a_{31} & a_{41} \\ a_{12} & a_{22} & a_{32} & a_{42} \end{pmatrix} + \begin{pmatrix} d_1^2 & 0 & 0 & 0 \\ 0 & d_2^2 & 0 & 0 \\ 0 & 0 & d_3^2 & 0 \\ 0 & 0 & 0 & d_4^2 \end{pmatrix}$$

$$\therefore \quad \boldsymbol{R} = \boldsymbol{A}\boldsymbol{A}^t + \boldsymbol{D} \tag{4.18}$$

$\boldsymbol{R} - \boldsymbol{D} = \boldsymbol{R}^*$ とおくと

$$\boldsymbol{R}^* = \boldsymbol{A}\boldsymbol{A}^t \tag{4.19}$$

なので，行列 $\boldsymbol{A}\boldsymbol{A}^t$ の非対角要素は相関行列 $\boldsymbol{R}$ 非対角要素に一致します[31]。また，$\boldsymbol{R}^*$ は**対称行列**ですね。

さて，相関行列 $\boldsymbol{R}$ はエクセルで簡単に計算できますが，(4.18) の右辺第 2 項，独自因子の行列 $\boldsymbol{D}$ が未知の数値なので，残念ながら (4.18) の連立方程式は解けません。そこで独自因子の部分を推定していくわけですが，勝手に推定するのでなく，次ぎに述べる**共通性**というものに注目して推定していくことになります。

4.2.3 共通性と独自性

(4.14) に戻って

$$a_{i1}^2 + a_{i2}^2 + d_i^2 = 1 \longrightarrow a_{i1}^2 + a_{i2}^2 = 1 - d_i^2 = h_i^2 \quad (i = 1, 2, 3, 4) \tag{4.20}$$

と変形します。左辺は共通因子によって説明できる部分なので「**共通性**」と呼んでいます。具体的にみていくと

$$\begin{cases} \text{・国語の共通性} \cdots & a_{11}^2 + a_{12}^2 = h_1^2 \\ \text{・英語の共通性} \cdots & a_{21}^2 + a_{22}^2 = h_2^2 \\ \text{・数学の共通性} \cdots & a_{31}^2 + a_{32}^2 = h_3^2 \\ \text{・理科の共通性} \cdots & a_{41}^2 + a_{42}^2 = h_4^2 \end{cases}$$

ところで「共通性」は一体何を意味しているのかということですが，それを調べるために (4.6) で独自因子が全くない場合を考えてみます。例えば変量「国語」をとり，それを z'_{i1} とすると

$$\begin{aligned} & z_{i1} = a_{i1}f_{i1} + a_{i2}f_{i2} + e_{i1} \\ \rightarrow \quad & z'_{i1} = a_{i1}f_{i1} + a_{i2}f_{i2} \quad (i = 1, 2, \cdots, n) \end{aligned}$$

z'_{i1} の分散を $V(z'_{i1})$ として，セクション 4.2.1 の手法に従って計算すると

$$\begin{aligned} V(z'_{i1}) &= \frac{1}{n-1}(z'^2_{11} + z'^2_{21} + \cdots + z'^2_{n1}) \\ &= \frac{1}{n-1}\left[(a_{11}f_{11} + a_{12}f_{12})^2 + (a_{21}f_{21} + a_{22}f_{22})^2 + \cdots + (a_{n1}f_{n1} + a_{n2}f_{n2})^2\right] \\ &= a_{11}^2 + a_{12}^2 \end{aligned}$$

[31] $\boldsymbol{D}$は対角行列。

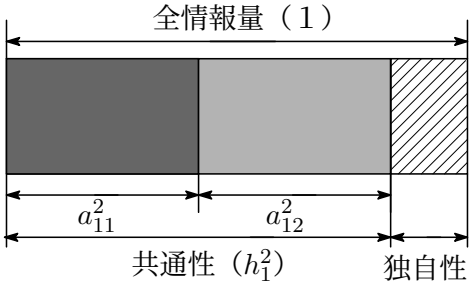
が得られます。つまり，独自因子のない変量「国語」の 分散 z'_{i1} は変量「国語」の“ 共通性 ”と等値ということです。他の変量も同様で，結果をまとめると次のようになります。

$$V(z'_{ik}) = a_{k1}^2 + a_{k2}^2 = h_k^2 \rightarrow \begin{cases} V(z'_{i1}) = a_{11}^2 + a_{12}^2 = h_1^2 & \rightarrow \text{国語の共通性} \\ V(z'_{i2}) = a_{21}^2 + a_{22}^2 = h_2^2 & \rightarrow \text{英語} \quad 〃 \\ V(z'_{i3}) = a_{31}^2 + a_{32}^2 = h_3^2 & \rightarrow \text{数学} \quad 〃 \\ V(z'_{i4}) = a_{41}^2 + a_{42}^2 = h_4^2 & \rightarrow \text{理科} \quad 〃 \end{cases} \tag{4.21}$$

（ただし，$i = 1, 2, \cdots, n; k = 1, 2, 3, 4$）

分散は変量の持つ情報量と考えられるので,「共通性」というのは共通因子だけで説明できる情報量ということになります。具体的には，変量「国語」のもつ全情報量を 1（標準化されているので分散は 1）とすると，h_1^2 は共通因子だけで説明できる情報の割合を表すことになります。これに対して，共通因子で説明されない部分，つまり，独自因子の分散である d_1^2 は国語に関する**独自性**と呼ばれます。(4.20) より $h_1^2 + d_1^2 = 1$ で，この関係を図示すると次のようになります。

《例：変量「国語」の全情報量の内訳》



共通性は一つひとつの変数 z_1, z_2, z_3, z_4 が持つ情報量のうち，共通する因子によって説明される割合を示しています。

また，各変量の共通性の和を h^2 で表し，

$$h^2 = h_1^2 + h_2^2 + h_3^2 + h_4^2 \tag{4.22}$$

を**総共通性**と呼びます。総共通性は，資料全体が共通因子によってどれだけ説明できるかを示しています。h^2 を具体的に書くと

$$\begin{aligned} h^2 &= (a_{11}^2 + a_{12}^2) + (a_{21}^2 + a_{22}^2) + (a_{31}^2 + a_{32}^2) + (a_{41}^2 + a_{42}^2) \\ &= (1 - d_1^2) + (1 - d_2^2) + (1 - d_3^2) + (1 - d_4^2) \\ &= 4 - (d_1^2 + d_2^2 + d_3^2 + d_4^2) \\ &= \text{変量の個数} - \text{独自因子の分散の合計} \end{aligned}$$

資料全体が因子分析モデル説明できる場合には，各変量の共通性は 1 になるので，総共通性 h^2 の最大値は変量の数です。そこで関係式が成立することが分かります。

$$\begin{aligned} \frac{h^2}{\text{変量の個数}} &= 1 - \frac{\text{独自因子の分散の合計}}{\text{変量の個数}} \\ &= \text{資料のもつ全情報のうち,共通因子で説明できる情報の割合} \end{aligned} \tag{4.23}$$

4.2.4 因子の寄与量

上で見てきたように，共通性は 2 つの共通因子が合わさって，全情報の内どれくらい説明するかを示す量でした。次ぎに，個々の共通因子が全情報のどれくらいの情報を説明しているかを調べて

みましょう。各共通因子に対する因子負荷量の平方和を次のようにとります。

$$C_1 = a_{11}^2 + a_{21}^2 + a_{31}^2 + a_{41}^2$$
$$C_2 = a_{12}^2 + a_{22}^2 + a_{32}^2 + a_{42}^2$$

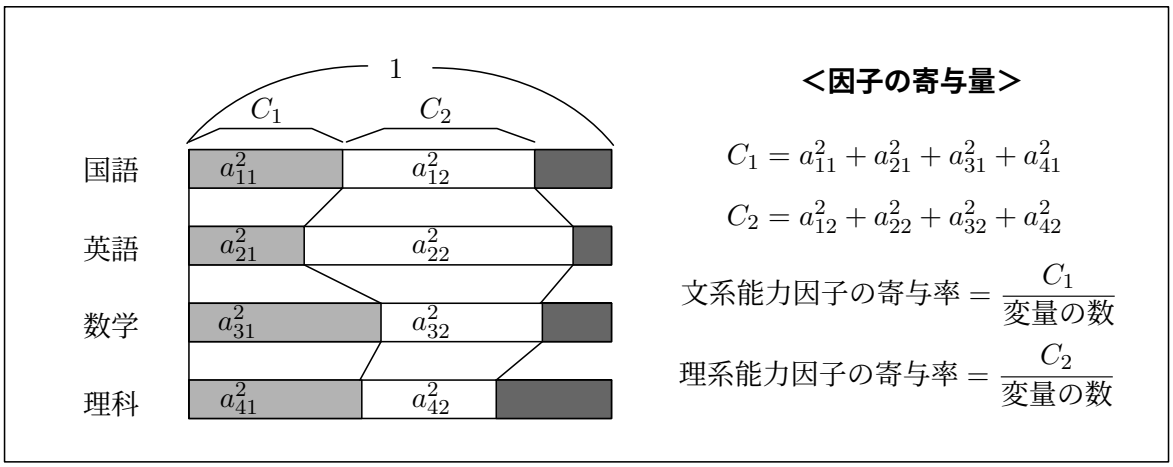
この 2 つの式の和をとると

$$\begin{aligned} C_1 + C_2 &= a_{11}^2 + a_{21}^2 + a_{31}^2 + a_{41}^2 + a_{12}^2 + a_{22}^2 + a_{32}^2 + a_{42}^2 \\ &= (a_{11}^2 + a_{12}^2) + (a_{21}^2 + a_{22}^2) + (a_{31}^2 + a_{32}^2) + (a_{41}^2 + a_{42}^2) \\ &= h_1^2 + h_2^2 + h_3^2 + h_4^2 \\ &= h^2 \quad (\longrightarrow \text{共通因子によって説明できる情報量}) \end{aligned} \tag{4.24}$$

となります。これから，次のことが言えます。

- C_1：共通因子によって説明できる情報量のうち，因子 1 から説明できる情報量
- C_2：共通因子によって説明できる情報量のうち，因子 2 から説明できる情報量

以上のことを図解すると次のようになります。



4.2.5 独自性の推定

さて，いよいよ独自性の推定に取り掛かることにします。まず (4.20) を眺めると，$h_i^2 + d_i^2 = 1$ なので，独自性 d_i^2 を推定することは共通性 h_i^2 を推定することと同じになります。このプロセスを簡単に纏めると，例えば国語の共通性 $h_1^2 (= a_{11}^2 + a_{12}^2)$ を推定することで独自性 $d_1^2 (= 1 - h_1^2)$ を推定するわけです。そしてこの推定値を使って a_{11} などを計算し，その結果を使って再び共通性 h_1^2（計算値）を求める。この作業を推定した共通性 h_1^2 の値と計算された共通性 h_1^2 の値が共に近い値に収束していくまで繰り返します。このような計算法[32]は，共通性の**反復推定**とか**反復解法**と呼ばれています。ここでは，この話にこれ以上立ち入ることはやめて，実際の計算は統計解析ソフトにやらすことにします。

さて，共通性の最初の推定値としては，重相関係数の 2 乗が使われます。重相関係数の 2 乗は，相関係数行列$\boldsymbol{R}$の逆行列$\boldsymbol{R}^{-1}$ の ii 要素を r^{ii} とした場合

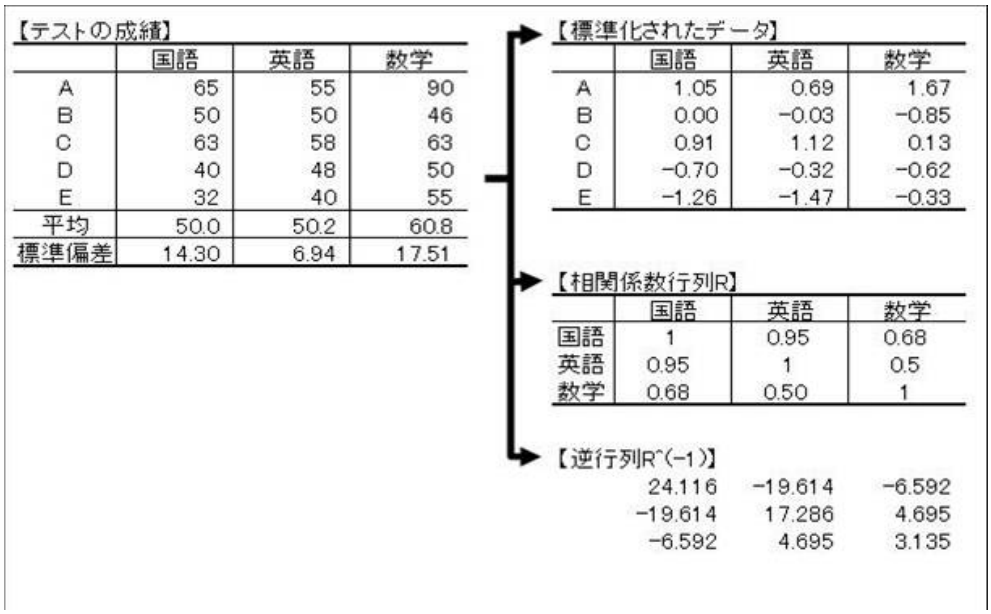
$$1 - \frac{1}{r^{ii}}$$

[32] 手計算では日が暮れるので，実際はパソコンで計算する。

で与えられるので，共通性の最初の推定値 h_i^2 は

$$h_i^2 = 1 - \frac{1}{r^{ii}} \tag{4.25}$$

となります。この方法を SMC 法 (squared multiple correlation method) といいます。具体例をやってみましょう。生徒 5 人のテストの成績が下表のようになったとします。この資料を標準化して，相関行列$\boldsymbol{R}$とその逆行列$\boldsymbol{R}^{-1}$ をエクセルで計算[33]しました。



【テストの成績】

	国語	英語	数学
A	65	55	90
B	50	50	46
C	63	58	63
D	40	48	50
E	32	40	55
平均	50.0	50.2	60.8
標準偏差	14.30	6.94	17.51

【標準化されたデータ】

	国語	英語	数学
A	1.05	0.69	1.67
B	0.00	−0.03	−0.85
C	0.91	1.12	0.13
D	−0.70	−0.32	−0.62
E	−1.26	−1.47	−0.33

【相関係数行列R】

	国語	英語	数学
国語	1	0.95	0.68
英語	0.95	1	0.5
数学	0.68	0.50	1

【逆行列R^(−1)】

24.116	−19.614	−6.592
−19.614	17.286	4.695
−6.592	4.695	3.135

(4.25) より

$$\begin{cases} h_1^2 = a_{11}^2 + a_{12}^2 = 1 - 1/24.116 = 0.96 \\ h_2^2 = a_{21}^2 + a_{22}^2 = 1 - 1/17.286 = 0.94 \\ h_3^2 = a_{31}^2 + a_{32}^2 = 1 - 1/3.135 = 0.68 \end{cases}$$

となり，また (4.16) と相関行列

$$\boldsymbol{R} = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{pmatrix}$$

より

$$\begin{cases} r_{12} = a_{11}a_{21} + a_{12}a_{22} = 0.95 \\ r_{13} = a_{11}a_{31} + a_{12}a_{32} = 0.68 \\ r_{23} = a_{21}a_{31} + a_{22}a_{32} = 0.50 \end{cases}$$

が得られます。

4.2.6 因子負荷量を求める（主因子法）

さて，最後に因子負荷量を求めていきますが，その代表的な手法としての**主因子法**について触れておきます。主因子法は寄与量の大きい因子を順次取り出していく方法です。因子負荷行列とその

[33] エクセル関数 minverse を使う。

転置行列は

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \end{pmatrix} = (\boldsymbol{a}_1, \boldsymbol{a}_2), \quad \boldsymbol{A}^t = \begin{pmatrix} a_{11} & a_{21} & a_{31} & a_{41} \\ a_{21} & a_{22} & a_{32} & a_{42} \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}_1^t \\ \boldsymbol{a}_2^t \end{pmatrix}$$

ただし，$\boldsymbol{a}_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \\ a_{41} \end{pmatrix}$, $\boldsymbol{a}_2 = \begin{pmatrix} a_{12} \\ a_{22} \\ a_{32} \\ a_{42} \end{pmatrix}$, $\boldsymbol{a}_1^t = (a_{11}, a_{21}, a_{31}, a_{41})$, $\boldsymbol{a}_2^t = (a_{12}, a_{22}, a_{32}, a_{42})$

で表されます。これから，

$$\boldsymbol{A}\boldsymbol{A}^t = \boldsymbol{a}_1\boldsymbol{a}_1^t + \boldsymbol{a}_2\boldsymbol{a}_2^t \tag{4.26}$$

と表すことができるので，(4.19) より

$$\boldsymbol{R}^* = \boldsymbol{A}\boldsymbol{A}^t = \boldsymbol{a}_1\boldsymbol{a}_1^t + \boldsymbol{a}_2\boldsymbol{a}_2^t \tag{4.27}$$

$\boldsymbol{R}^*$ は 4 行 4 列の対称行列なので，4 個の固有値（λ_1, λ_2, λ_3, λ_4）と互いに直交する 4 つの固有ベクトル $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, $\boldsymbol{x}_4$ をもちます。これらを使うと $\boldsymbol{R}^*$ は

$$\boldsymbol{R}^* = \lambda_1\boldsymbol{x}_1\boldsymbol{x}_1^t + \lambda_2\boldsymbol{x}_2\boldsymbol{x}_2^t + \lambda_3\boldsymbol{x}_3\boldsymbol{x}_3^t + \lambda_4\boldsymbol{x}_4\boldsymbol{x}_4^t \tag{4.28}$$

と表すことができます[34]。ここで固有値 λ_i は

$$\lambda_1 \geqq \lambda_2 \geqq \lambda_3 \geqq \lambda_4$$

の順になっているとします[35]。(4.27) と (4.28) より

$$\boldsymbol{a}_1\boldsymbol{a}_1^t + \boldsymbol{a}_2\boldsymbol{a}_2^t = \lambda_1\boldsymbol{x}_1\boldsymbol{x}_1^t + \lambda_2\boldsymbol{x}_2\boldsymbol{x}_2^t + \lambda_3\boldsymbol{x}_3\boldsymbol{x}_3^t + \lambda_4\boldsymbol{x}_4\boldsymbol{x}_4^t \tag{4.29}$$

さて，λ_1, λ_2 が正の値で，λ_3, λ_4 に比べて大きいとすると，(4.29) の右辺第 3，4 項は無視して

$$\boldsymbol{a}_1 = \sqrt{\lambda_1}\boldsymbol{x}_1, \quad \boldsymbol{a}_2 = \sqrt{\lambda_2}\boldsymbol{x}_2 \tag{4.30}$$

と近似してもいいでしょう。これからベクトル $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ が近似的に求められ，因子負荷行列 $\boldsymbol{A}$ の各成分が求めらるわけです。こうして因子負荷量を求める方法を**主因子法**といいます。

今，独自因子の部分が無視できるとした場合，

$$\boldsymbol{R}^* = \boldsymbol{R} \tag{4.31}$$

で $\boldsymbol{R}^*$の固有値，固有ベクトルは相関行列$\boldsymbol{R}$の固有値，固有ベクトルに一致します。ところで，主成分分析のセクションの (3.32) で見たように，標準化したデータの主成分分析は相関行列の固有値問題に帰しました。つまり，(4.28) の $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ がそれぞれ第 1 主成分，第 2 主成分 ということになります。したがって (4.30) の因子負荷量の成分ベクトル $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ は，その大きさを除いて第 1 主成分，第 2 主成分と同一ということになります。

[34] 以下に証明の要点だけ。対称行列を$\boldsymbol{H}$とし，その固有値を h_i，固有ベクトルを $|h_i\rangle$ とすると，$\boldsymbol{G}|h_i\rangle = h_i|h_i\rangle$．固有ベクトルの完備性条件より $\sum_i |h_i\rangle\langle h_i| = \boldsymbol{I}$．ただし$\boldsymbol{I}$は単位行列．これから $\boldsymbol{G}\sum_i |h_i\rangle\langle h_i| = \boldsymbol{G} = \sum_i h_i|h_i\rangle\langle h_i|$

[35] 負の固有値は寄与量が負となり意味を持たないので正の固有値だけを採用する。

4.2.7　因子の解釈（バリマックス法）

以上のプロセスで因子負荷行列$\boldsymbol{A}$を求めることができました。ここで (4.14), (4.16) の関係式を再掲すると

$$(4.14)\quad \begin{cases} a_{11}^2 + a_{12}^2 + d_1^2 = 1 \\ a_{21}^2 + a_{22}^2 + d_2^2 = 1 \\ a_{31}^2 + a_{32}^2 + d_3^2 = 1 \\ a_{41}^2 + a_{42}^2 + d_4^2 = 1 \end{cases}$$

$$(4.16)\quad \begin{cases} r_{12} = a_{11}a_{21} + a_{12}a_{22}, \quad r_{13} = a_{11}a_{31} + a_{12}a_{32} \\ r_{14} = a_{11}a_{41} + a_{12}a_{42}, \quad r_{23} = a_{21}a_{31} + a_{22}a_{32} \\ r_{24} = a_{21}a_{41} + a_{22}a_{42}, \quad r_{34} = a_{31}a_{41} + a_{32}a_{42} \end{cases}$$

この式の構造を調べてみます。$\boldsymbol{a} = (a_{11},\, a_{12})$, $\boldsymbol{b} = (a_{11},\, a_{12})$ とおくと (4.14) の第 1 式は

$$\boldsymbol{a} \cdot \boldsymbol{b} + d_1^2 = 1 \tag{4.32}$$

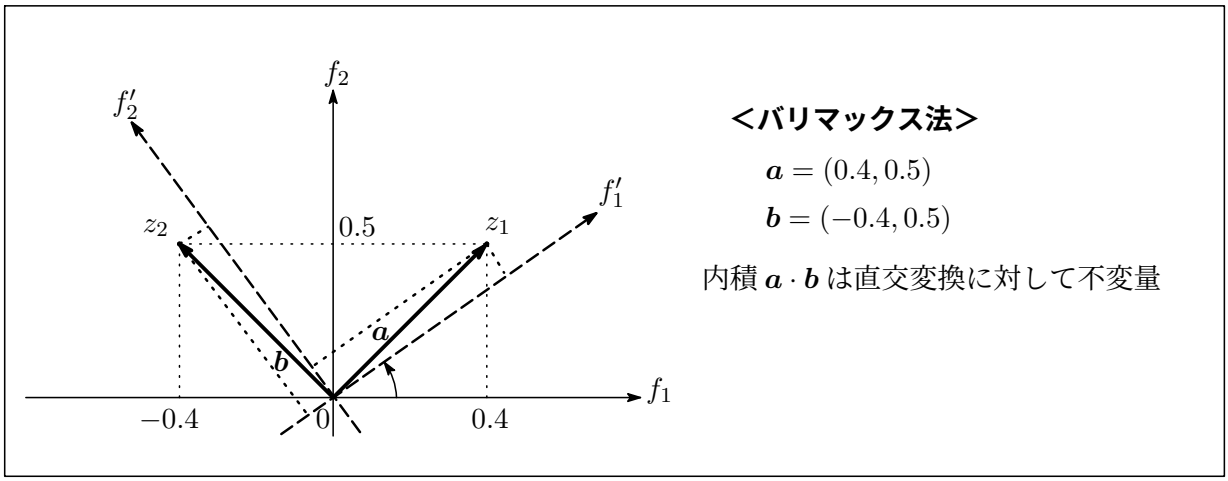
また，$\boldsymbol{a} = (a_{11},\, a_{12})$, $\boldsymbol{b} = (a_{21},\, a_{22})$ とおくと (4.16) の第 1 式は

$$\boldsymbol{a} \cdot \boldsymbol{b} = r_{12} \tag{4.33}$$

と，因子負荷量の基本式はベクトルの内積で表わすことができます（←他の式も同様に内積で表せる）。内積はスカラー量で座標軸の回転に対して不変量です。ということで，(4.14)，(4.16) で得られた方程式の解は，座標軸を回転しても元の方程式の解であることには変わりがありません。つまり解は回転の不定性があるということですね。そこで，求めた解を解釈しやすいように，いくつかの変量の因子負荷量の絶対は大きく，残りの変量の因子負荷量は 0 に近くなるように座標軸の直交変換をしてやります。この方法を**バリマックス法**といいます[36]。具体的に図で示すと，いま，変量 z_1, z_2 があって，

$$z_1 = 0.4f_1 + 0.5f_2, \quad z_2 = -0.4f_1 + 0.5f_2$$

で表されるとします。この因子負荷量のベクトルは $\boldsymbol{a} = (0.4,\, 0.5)$, $\boldsymbol{b} = (-0.4,\, 0.5)$ で表せるので，縦軸を因子 f_1，横軸を因子 f_2 と平面座標軸を，変量 z_1 を表す点にできるだけ近付くように新たな因子軸 f_1' を，変量 z_2 を表す点にできるだけ近付くように新たな因子軸 f_2' を描きます（ただし f_1', f_2' は直交座標系）。



[36] 主成分分析でやった手法とよく似ていますね！

4.3　因子分析の例

計算は「エクセル統計」に添付されているエクセル・アドインソフト Mulcel を使いました。

【例－１】
コンビニ 10 店舗のアンケート結果（「図解で分かる多変量解析」4 章 7 の項参照）を因子分析します。共通因子として次の 2 つの共通因子を仮定します。

「ハード因子」: 立地場所等，地理的・物理的条件要因による因子

「ソフト因子」: 客を呼び寄せるための展示工夫等要因による因子

さて，コンビニ 10 店舗のアンケートでの評点は下表の通りでした。

店名	品揃え	雰囲気	親近感	広々感
1 号店	20	20	30	70
2 号店	15	25	25	15
3 号店	60	70	60	30
4 号店	70	50	60	40
5 号店	30	35	55	80
6 号店	50	85	85	80
7 号店	80	65	70	30
8 号店	30	65	75	80
9 号店	60	60	50	40
10 号店	25	25	50	25

■ Mulcel による因子分析の結果
統計解析ソフト Mulcel を使った因子分析の結果をこの項の最後に載せます。尚，表内の a_1, a_2 はそれぞれ**ソフト因子**，**ハード因子**です。

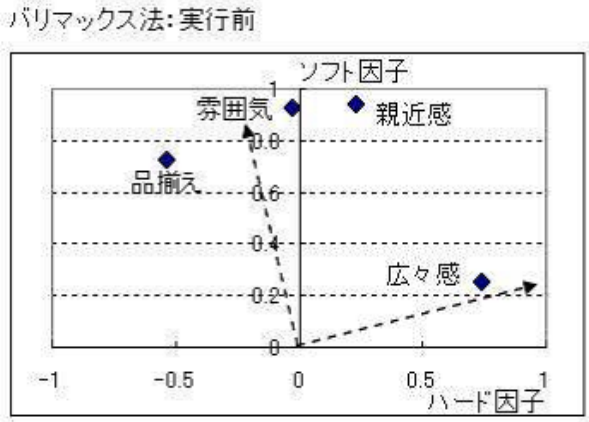
(1) 共通性の推定値　「SMC 法」の「非反復解法」と「反復解法」のところを見てみると共通性の値の変化（収束状況）がわかります。eps は収束判定条件で $|r^*_{ij} - \sum_{k=1}^{m} a^2_{ik}| < \varepsilon(= 0.00001)$。

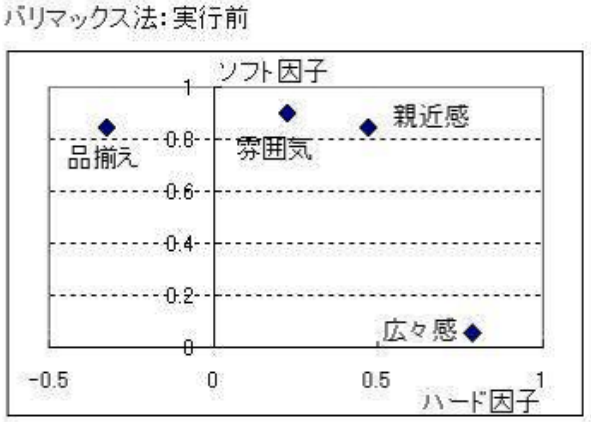
	h^2（初期値）	h^2（反復推定値）
品揃え	0.6903	0.8115
雰囲気	0.8561	0.8530
親近感	0.8675	0.9311
広々感	0.5098	0.6198

(2) 因子負荷量　バリマックス法の実行前後のソフト因子，ハード因子は

	【前】		【後】	
	a_1（ソフト因子）	a_2（ハード因子）	a_1（ソフト因子）	a_1（ソフト因子）
品揃え	0.7255	−0.5340	0.7255	−0.5340
雰囲気	0.9233	−0.0225	0.9233	−0.0225
親近感	0.9369	0.2311	0.9369	0.2311
広々感	0.2577	0.7440	0.2577	0.7439

これを縦軸をソフト因子，横軸をハード因子とした座標上にプロットすると



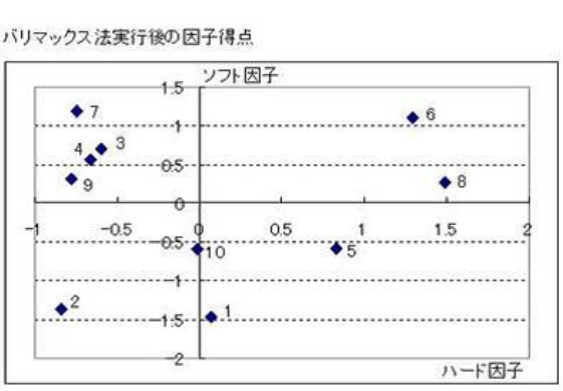


となります。特に「広々感」はコンビニの敷地の広さ，つまりハード因子と深く関連していますが，バリマックス法実行前ではその特長はやや不明瞭です。バリマックス法実行後ではハード因子軸に近付き，その特長を明確にしています。同時に，「品揃え」，「雰囲気」「親近感」はすべて上方に追いやられ，ソフト因子の特長をいっそう際立たせることが分かります。

(3) 各店の因子得点　各点の因子得点は下表のように得られます。

店番号	f1	f2
1	-1.4669	0.0695
2	-1.3687	-0.8407
3	0.7049	-0.5970
4	0.5565	-0.6616
5	-0.5858	0.8292
6	1.0917	1.2928
7	1.1887	-0.7440
8	0.2653	1.4849
9	0.3075	-0.7791
10	-0.6932	-0.054

これを座標上にプロットすると右図のようになります。このグラフからコンビニ各店舗のソフト因子，ハード因子保有量が一目瞭然に分かります。例えば 1 号店は，ハード因子は全店の平均値に近いですが，ソフト因子は足りないようです。店舗の飾りやサービスの拡充等の経営努力が必要なようです。



(4) 寄与率　バリマックス法実行後の寄与率は

	a_1	a_2	h^2
寄与量	2.2252	0.9903	3.2155
寄与率	0.5563	0.2476	0.8039

で，2つの因子で資料のもち全情報の約80％を説明しています。かなり精度が高いといえますね！

因子分析

データ数　10

変量	平均	不偏分散	標準偏差	標準誤差
品揃え	44	521.1111	22.82786	7.218803
雰囲気	50	505.5556	22.48456	7.110243
親近感	56	348.8889	18.67857	5.906682
広々感	49	660	25.69047	8.124038

相関行列

	品揃え	雰囲気	親近感	広々感
品揃え	1	0.681896	0.556348	−0.2103
雰囲気	0.681896	1	0.85983	0.221207
親近感	0.556348	0.85983	1	0.413314
広々感	−0.2103	0.221207	0.413314	1

相関行列の固有値

2.458439	1.203699	0.222393	0.115469

対角要素をSMCでおきかえた相関行列の固有値

2.205807	0.718055	−0.08972	−0.16748

SMC法

非反復解法
因子負荷量と共通性

	a1	a2	h^2
品揃え	0.681626	−0.47514	0.690368
雰囲気	0.924649	−0.0339	0.856125
親近感	0.907917	0.207826	0.867506
広々感	0.248804	0.669299	0.509864
寄与量	2.205807	0.718055	2.923863
寄与率	0.551452	0.179514	0.730966

反復解法　eps=0.00001　最大反復回数=　100
因子負荷量と共通性　反復回数=　28

	a1	a2	h^2
品揃え	0.725524	−0.53398	0.811525
雰囲気	0.923325	−0.02247	0.853034
親近感	0.936868	0.231063	0.931112
広々感	0.25768	0.743924	0.619822
寄与量	2.323036	0.892457	3.215494
寄与率	0.580759	0.223114	0.803873

バリマックス回転　eps=0.00001
反復解法
初期バリマックス基準値　4.000428

因子負荷量と共通性

	a1	a2	h^2
品揃え	0.725524	−0.53398	0.811525
雰囲気	0.923325	−0.02247	0.853034
親近感	0.936868	0.231063	0.931112
広々感	0.25768	0.743924	0.619822
寄与量	2.323036	0.892457	3.215494
寄与率	0.580759	0.223114	0.803873

回転後のバリマックス基準値　4.508992

因子負荷量と共通性

	a1	a2	h^2
品揃え	0.839928	−0.32565	0.811525
雰囲気	0.897065	0.219791	0.853034
親近感	0.843828	0.468046	0.931112
広々感	0.054148	0.785423	0.619822
寄与量	2.225183	0.990311	3.215494
寄与率	0.556296	0.247578	0.803873

因子得点の推定

	f1	f2
1	−1.46686	0.06949
2	−1.36867	−0.84074
3	0.70494	−0.59704
4	0.556482	−0.66158
5	−0.58578	0.829173
6	1.091689	1.292839
7	1.188677	−0.74398
8	0.265262	1.484941
9	0.307467	−0.77914
10	−0.69321	−0.05396

5　正準相関分析

——- いつになるやら ————

6　判別分析

by $\mathscr{KENZOU}$

（了）