統計学における FAQ

青木繁伸

2013 年 5 月 16 日

1 重回帰分析において

1.1 重回帰分析における誤解

1.1.1 「順序尺度データは重回帰分析には使えない」というのは誤り

10 年くらい前までは，順序尺度データや名義尺度データなど，いわゆるカテゴリーデータは重回帰分析や判別分析には使えないから，数量化 I 類や数量化 II 類を使いなさいと良く言われていた。SPSS に数量化 I 類，II 類，III 類がないのは困るということで，アドインプログラムが提供されたりした。

今は，ダミー変数を使って普通に重回帰分析や判別分析が行われている。

有名なアイリスデータ iris は 4 つの連続変数 Sepal.Length，Sepal.Width，Petal.Length，Petal.Width と 3 つのカテゴリーを持つカテゴリー変数 Species からなる。

```
    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1            5.1         3.5          1.4         0.2  setosa
2            4.9         3.0          1.4         0.2  setosa
3            4.7         3.2          1.3         0.2  setosa
4            4.6         3.1          1.5         0.2  setosa
5            5.0         3.6          1.4         0.2  setosa
6            5.4         3.9          1.7         0.4  setosa
```

R で一般的なデータ構造 data.frame は，データ入力時・作成時に文字列データは Factor にされる。Factor を多変量解析の対象とすると特別なことは何もしないでも適正に扱われる。

Sepal.Width，Petal.Length，Petal.Width と Species を説明変数として Sepal.Length を予測する重回帰分析の結果を以下に示す。

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.17127    0.27979   7.760 1.43e-12
Sepal.Width        0.49589    0.08607   5.761 4.87e-08
Petal.Length       0.82924    0.06853  12.101  < 2e-16
Petal.Width       -0.31516    0.15120  -2.084  0.03889
Speciesversicolor -0.72356    0.24017  -3.013  0.00306
Speciesvirginica  -1.02350    0.33373  -3.067  0.00258

Multiple R-squared: 0.8673,        Adjusted R-squared: 0.8627
F-statistic: 188.251 on 5 and 144 DF,  p-value: < 2.2e-16
```

この分析結果を見ると，Species は，Speciesversicolor と Speciesvirginica の 2 つで扱われている。Estimate の列が偏回帰係数であるが，分析結果に表れない Speciessetosa の偏回帰係数は 0 なのである。Speciessetosa に比べて Speciesversicolor は −0.72356 （つまり 0.72356 小さい），おなじく，Speciesvirginica は −1.02350 ということである。

このような特別な機能を持たない統計解析ソフトを用いる場合には，カテゴリー変数をダミー変数に変換する必要がある。

ダミー変数は 0/1 のような二値変数で表されるもので，一つの変数では 2 種類の状態しか表現できない。k 種類のカテゴリーを持つカテゴリー変数が持つ情報を完全に表すには $k-1$ 個のダミー変数が必要である。

表 1 ダミー変数での表現

	ダミー変数 1	ダミー変数 2
setosa	0	0
versicolor	1	0
virginica	0	1

versicolor，verginica を表す 2 つのダミー変数を明示的に作って，重回帰分析をしてみよう。

```
    Sepal.Length Sepal.Width Petal.Length Petal.Width versicolor virginica
1            5.1         3.5          1.4         0.2          0         0
2            4.9         3.0          1.4         0.2          0         0
3            4.7         3.2          1.3         0.2          0         0
51           7.0         3.2          4.7         1.4          1         0
52           6.4         3.2          4.5         1.5          1         0
53           6.9         3.1          4.9         1.5          1         0
101          6.3         3.3          6.0         2.5          0         1
102          5.8         2.7          5.1         1.9          0         1
103          7.1         3.0          5.9         2.1          0         1
```


分析結果は以下のようになる。

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.17127    0.27979   7.760 1.43e-12
Sepal.Width   0.49589    0.08607   5.761 4.87e-08
Petal.Length  0.82924    0.06853  12.101  < 2e-16
Petal.Width  -0.31516    0.15120  -2.084  0.03889
versicolor   -0.72356    0.24017  -3.013  0.00306
virginica    -1.02350    0.33373  -3.067  0.00258

Multiple R-squared: 0.8673,        Adjusted R-squared: 0.8627
F-statistic: 188.251 on 5 and 144 DF,  p-value: < 2.2e-16
```

以上のように，Factor を使った重回帰分析と同じ結果が得られることが確かめられた。

1.1.2 「回帰の分散分析の P 値が有意ならば優れたモデルである」というのは誤り

検定についての誤解の一つに，「P 値が小さいことは，主張したいことが正しいという証拠になる」というのがある。しかし，あらゆる検定において，P 値は検定統計量が帰無仮説のもとで起こりにくいことを示すが，サンプルサイズにも依存するということである。回帰の分散分析の帰無仮説は「回帰により説明される分散は誤差分散なみである」ということであり，P 値が小さいということは「回帰により説明される分散は誤差分散に比べて十分に大きい」ということである。しかし，重回帰式が使えるかどうかは決定係数を見るべきである。サンプルサイズがちょっと大きければ，決定係数が小さくても回帰の分散分析の帰無仮説は棄却される。

表 2 サンプルサイズが大きい場合の重回帰分析の結果

	偏回帰係数	標準誤差	t 値	P 値	標準化偏回帰係数	VIF
x1	0.190	0.055	3.459	< 0.001	0.190	1.190
x2	0.300	0.058	5.148	< 0.001	0.300	1.333
x3	0.124	0.062	1.987	0.048	0.124	1.524
定数項	0.000	0.050	0.000	1.000		

重相関係数 $R = 0.470$
重相関係数の二乗（決定係数）$R^2 = 0.220$
自由度調整済み重相関係数の二乗 $= 0.213$
回帰の分散分析：F 値 $(3, 306) = 28.849$，P 値 < 0.001

独立変数 x_1，x_2，x_3 の偏回帰係数の検定の P 値は全て 0.05 より小さく，帰無仮説は棄却される。さらに回帰の分散分析の結果は $F(3, 306)$=28.849 であり，P 値 $< 2.2e{-}16$ で，帰無仮説は棄却される。しかし，決定係数 R^2 は 0.220 であり，図 1 に示すように，とても有効な予測ができるようなものではないことがわかる。

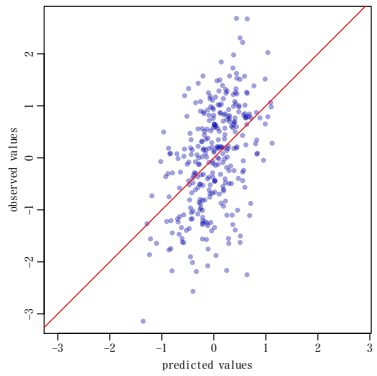

図 1 予測値と実測値

1.1.3 「独立変数間の相関係数に 0.8 を超えるものがあると多重共線性が生じる」というのは誤り

多重共線性が生じるかどうかは，分析に使用する全ての変数の相関関係によるので，実際に計算してみなければわからない。

- 例 1

  表 3 のような相関係数行列を持つデータを考える。相関係数は 0.8 以上のものばかりである。

表 3 相関係数行列

	y	x1	x2	x3
y	1.00	0.89	0.85	0.88
x1	0.89	1.00	0.85	0.87
x2	0.85	0.85	1.00	0.83
x3	0.88	0.87	0.83	1.00

重回帰分析の結果は，以下のようになる。回帰係数の符号，大きさ，VIF も特に問題はない。

表 4　相関の高い変数を用いた重回帰分析の結果

	偏回帰係数	標準誤差	t 値	P 値	標準化偏回帰係数	VIF
x1	0.381	0.144	2.638	0.012	0.381	5.356
x2	0.232	0.125	1.854	0.071	0.232	4.041
x3	0.353	0.136	2.604	0.013	0.353	4.745
定数項	0.000	0.062	0.000	1.000		

重相関係数 $R = 0.919$
重相関係数の二乗（決定係数）$R^2 = 0.845$
自由度調整済み重相関係数の二乗 = 0.833
回帰の分散分析：F 値 $(3, 40) = 72.483$，P 値 < 0.001

- 例 2

以下のような相関係数行列を持つデータを考える。0.8 以上の相関係数は 1 個しかない。

表 5　相関係数行列

	y	x1	x2	x3
y	1.00	0.51	0.56	0.55
x1	0.51	1.00	0.95	0.52
x2	0.56	0.95	1.00	0.58
x3	0.55	0.52	0.58	1.00

重回帰分析の結果は，以下のようになる。VIF を見ると，独立変数 x_1 か x_2 に問題がありそうだ。

表 6　相関の高い変数を用いた重回帰分析の結果

	偏回帰係数	標準誤差	t 値	P 値	標準化偏回帰係数	VIF
x1	−0.086	0.400	−0.216	0.830	−0.086	10.457
x2	0.437	0.421	1.039	0.305	0.437	11.586
x3	0.344	0.154	2.234	0.031	0.344	1.549
定数項	−0.000	0.122	−0.000	1.000		

重相関係数 $R = 0.624$
重相関係数の二乗（決定係数）$R^2 = 0.389$
自由度調整済み重相関係数の二乗 = 0.343
回帰の分散分析：F 値 $(3, 40) = 8.497$，P 値 < 0.001

1.2　多重共線性の回避策

独立変数 x_4 が従属変数 y に与える影響を知りたい。ほかの独立変数 x_1 と x_2 の間に多重共線性があるが，これはどちらも説明変数に含めたい。x_1 と x_4，x_2 と x_4 の相関は低く VIF も小さい。

このような場合にどうしたらよいかという質問に対して，シミュレーションに基づく解答を提示する。

変数間の相関係数行列は表 7 のようになっていると仮定しよう。

表 7　相関係数行列

	x_1	x_2	x_3	x_4	y
x_1	1.00	0.80	0.30	0.20	0.30
x_2	0.80	1.00	0.30	0.20	0.25
x_3	0.30	0.30	1.00	0.30	0.30
x_4	0.20	0.20	0.30	1.00	0.40
y	0.30	0.25	0.30	0.40	1.00

まず，表 7 のような相関係数行列になるように，多変量正規分布にしたがうシミュレーションデータを作成する（$n = 500$）。

先頭の 10 行は表 8 のようになる（小数点以下 4 桁以降は表示していない）。

表 8　データ行列

	x_1	x_2	x_3	x_4	y
1	35.871	36.708	73.362	66.540	33.123
2	37.626	29.151	82.969	76.820	33.955
3	59.216	55.500	77.283	86.957	36.323
4	63.243	48.876	84.256	81.253	36.780
5	48.795	39.846	64.629	98.008	39.801
6	48.312	51.324	89.104	99.748	39.748
7	44.331	31.496	74.814	86.109	37.237
8	68.218	64.982	76.578	86.126	36.145
9	40.228	44.737	81.746	71.228	33.558
10	57.571	60.432	80.294	80.866	35.988

それぞれの基礎統計量は表 9 の通りである。

表 9　各変数の平均値と不偏分散および標準偏差

	平均値	不偏分散	標準偏差
x_1	50.0	100.0	10.0
x_2	46.0	81.0	9.0
x_3	76.0	36.0	6.0
x_4	82.0	64.0	8.0
y	36.0	4.0	2.0

散布図は図 2 のようになり，相関係数行列は正確に表 7 の通りになる。

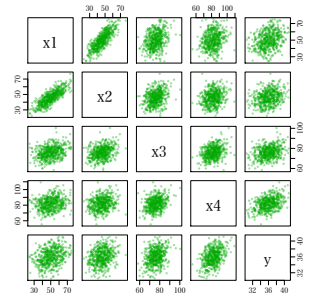

図 2　散布図

x_1 と x_2 をそのまま使用すると，結果は表 10 のようになる。x_1 と x_2 の VIF（分散拡大要因）は大きなものではないが，y との相関係数が 0.3，0.25 とほぼ同じなのに標準化偏回帰係数の絶対値は 7 倍も違い，符号は逆になっている。この状態が多重共線性か否かについては即断はできないが[*1]，解釈に困るのは確かであろう。

表 10　x_1 と x_2 をそのまま使用する場合

	偏回帰係数	標準誤差	t 値	P 値	標準化偏回帰係数	VIF
x1	0.044	0.013	3.287	0.001	0.218	2.814
x2	−0.007	0.015	−0.490	0.625	−0.032	2.814
x3	0.050	0.014	3.466	< 0.001	0.149	1.186
x4	0.080	0.010	7.631	< 0.001	0.318	1.117
定数項	23.857	1.137	20.986	< 0.001		

重相関係数 $R = 0.479$
重相関係数の二乗（決定係数）$R^2 = 0.229$
自由度調整済み重相関係数の二乗 = 0.223
回帰の分散分析：F 値 $(4, 495) = 36.792$，P 値 < 0.001

[*1] x_2, x_3, x_4 を制御したときの x_1 と y の偏相関係数は 0.146 であるが，x_1, x_3, x_4 を制御したときの x_2 と y の偏相関係数は −0.022 なので，抑制変数の可能性はある。

x_1 と x_2 の相関が高いために困った事態が生じるのであるから，解決策は「どちらかだけを使う」ことであるのは明らかである。表 11，12 に示すように x_1 と x_2 の標準化偏回帰係数の値は，それらと y の相関係数に見合ったものであること，また，x_3，x_4 の標準化偏回帰係数はほぼ同じになることがわかる。

表 11　x_1 だけを使用する場合

	偏回帰係数	標準誤差	t 値	P 値	標準化偏回帰係数	VIF
x1	0.038	0.008	4.621	< 0.001	0.192	1.115
x3	0.049	0.014	3.438	< 0.001	0.147	1.176
x4	0.079	0.010	7.622	< 0.001	0.317	1.115
定数項	23.844	1.136	20.996	< 0.001		

重相関係数 $R = 0.478$
重相関係数の二乗（決定係数）$R^2 = 0.229$
自由度調整済み重相関係数の二乗 = 0.224
回帰の分散分析：F 値 $(3, 496) = 49.051$，P 値 < 0.001

表 12　x_2 だけを使用する場合

	偏回帰係数	標準誤差	t 値	P 値	標準化偏回帰係数	VIF
x2	0.030	0.009	3.247	0.001	0.137	1.115
x3	0.054	0.014	3.743	< 0.001	0.162	1.176
x4	0.081	0.011	7.703	< 0.001	0.324	1.115
定数項	23.860	1.148	20.784	< 0.001		

重相関係数 $R = 0.461$
重相関係数の二乗（決定係数）$R^2 = 0.212$
自由度調整済み重相関係数の二乗 = 0.208
回帰の分散分析：F 値 $(3, 496) = 44.573$，P 値 < 0.001

もう一つの解決法は，x_1 と x_2 の相関が高いということはどちらの変数もよく似たものを測定しているのだから両方足してしまう，つまり，$x_1 + x_2$ を新たな独立変数として使うということである。この方法は，2 つの変数が平均値も分散もほとんど同じなら単純に和をとることにも違和感がないかもしれないが通常はそのようなことは期待できない。そこで，2 つの変数をそれぞれ標準化すれば平均値と分散は同じになるのだから足し算をすることができる。そして，単純和よりも望ましい方法は重み付けの和をとることであろう。

さて，解決法が見えてきた。2 つの変数だけで主成分分析を行い，主成分得点を独立変数とすることである。2 つの変数の主成分分析からは 2 つの主成分得点が得られる。もとの 2 つの変数が持つ情報は 2 つの主成分が持つ情報と全く同じである。そして，2 つの主成分得点の相関は 0 である。

表 13　主成分分析による回転行列（重み）

	第 1 主成分	第 2 主成分
x_1	0.70711	−0.70711
x_2	0.70711	0.70711

表 14　主成分得点

	第 1 主成分	第 2 主成分
1	−1.72912	0.26898
2	−2.19875	−0.44874
3	1.39806	0.09473
4	1.16242	−0.71043
5	−0.56875	−0.39832
6	0.29892	0.53771
7	−1.54038	−0.73872
8	2.77957	0.20311
9	−0.79016	0.59176
10	1.66920	0.59851

表 15　各変数の平均値と不偏分散および標準偏差

	平均値	不偏分散	標準偏差
第 1 主成分	0.00000	1.80000	1.34164
第 2 主成分	−0.00000	0.20000	0.44721

表 16　主成分得点ともとの変数の相関

	PC_1	PC_2	x_3	x_4	y
PC_1	1.00000	−0.00000	0.31623	0.21082	0.28988
PC_2	−0.00000	1.00000	−0.00000	−0.00000	−0.07906
x_3	0.31623	−0.00000	1.00000	0.30000	0.30000
x_4	0.21082	−0.00000	0.30000	1.00000	0.40000
y	0.28988	−0.07906	0.30000	0.40000	1.00000

では，x_1 と x_2 の代わりに 2 つの主成分得点を用いて重回帰分析を行ってみよう。結果を表 17 に示す。x_3，x_4 に対する標準化偏回帰係数は，表 10，11，12 の標準化偏回帰係数とほぼ同じになっている。

表 17　x_1 と x_2 の主成分得点を使用する場合

	偏回帰係数	標準誤差	t 値	P 値	標準化偏回帰係数	VIF
PC1	0.262	0.063	4.188	< 0.001	0.176	1.130
PC2	−0.354	0.176	−2.003	0.046	−0.079	1.000
x3	0.050	0.014	3.466	< 0.001	0.149	1.186
x4	0.080	0.010	7.631	< 0.001	0.318	1.117
定数項	25.702	1.206	21.312	< 0.001		

重相関係数 $R = 0.479$
重相関係数の二乗（決定係数）$R^2 = 0.229$
自由度調整済み重相関係数の二乗 = 0.223
回帰の分散分析：F 値 $(4, 495) = 36.792$，P 値 < 0.001

なお，このように独立変数の値をそのまま使うのではなく主成分分析によって得られる主成分得点を使って重回帰分析をするのは主成分回帰（Partial Compnent Regression; PCR）として確立されているものである。しかし，全ての独立変数を主成分分析する必要はなく，必要なもののみを主成分分析し主成分回帰を行うというのが自然なのである。

1.3　主成分回帰

重回帰分析をする際，説明変数の中に互いに相関が高い変数が含まれる場合，通常の最小 2 乗法 では回帰係数の推定精度が悪くなるという問題（多重共線性）がある。このような場合の対処法として，以下のものがある。

- PCR 回帰 (Principal Component Regression：主成分回帰)
  説明変数に対して主成分分析を行い，その主成分得点を使って従属変数を予測する
- PLS 回帰 (Partial Least Squares Regression)
  説明変数から潜在変数を計算し，そのスコアで従属変数を予測する（潜在変数が主成分で定義されると PCR と同じ）
- リッジ回帰 (Ridge Rgression：RR)
  正規方程式の係数行列の対角成分を調整して重回帰分析を行う

主成分回帰は，どのような場合にも適用できるものではないことに注意が必要である。具体的には，以下のような場合に適用する。3 番目は重要（必然的に 2 番目も）。

- 相関の高い複数の独立変数を使うとき（多重共線性があるとき）
- 独立変数の測定単位は同じでなくてはならない
- 標準化しないで分析する（分散共分散行列を対象にする）

1.3.1　主成分回帰と直線回帰の違い

重回帰分析では，回帰超平面上の予測値と実測値の差の二乗和を最小にするという考え方をとった。主成分回帰では，回帰超平面に下ろした垂線の長さの二乗和を最小にするという考え方をする。

両者の解析例の違いを示すためのテストデータを作成する。

図 3 で，赤が直線回帰，青が主成分回帰である。常に，主成分回帰直線の傾きの方が大きい。

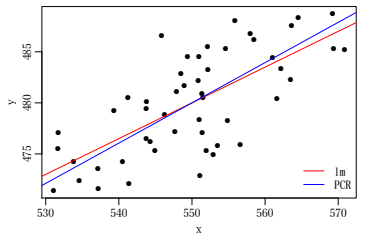

図 3　直線回帰と主成分回帰の違い

1.3.2　独立変数の中に相関の高いものがある場合の実例

iris データセットを使って説明する。

相関係数は表 18 のようになる。

Sepal.Length を従属変数として，残りの 3 変数を独立変数として重回帰分析する場合を考える。独立変数の Petal.Length と Petal.Width の相関が高いのが問題となる。

表 18　相関係数行列

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.000	−0.118	0.872	0.818
Sepal.Width	−0.118	1.000	−0.428	−0.366
Petal.Length	0.872	−0.428	1.000	0.963
Petal.Width	0.818	−0.366	0.963	1.000

重回帰分析の結果は表 19 のようになるが，この結果はそのまま受け入れることはできない。

表 19　重回帰分析の結果

	偏回帰係数	標準誤差	t 値	P 値	標準化偏回帰係数	VIF
Sepal.Width	0.651	0.067	9.765	< 0.001	0.343	1.271
Petal.Length	0.709	0.057	12.502	< 0.001	1.512	15.098
Petal.Width	−0.556	0.128	−4.363	< 0.001	−0.512	14.234
定数項	1.856	0.251	7.401	< 0.001		

重相関係数 $R = 0.927$
重相関係数の二乗（決定係数）$R^2 = 0.859$
自由度調整済み重相関係数の二乗 = 0.856
回帰の分散分析：F 値 $(3, 146) = 295.539$，P 値 < 0.001

Petal.Length, Petal.Width の VIF はそれぞれ 15.098，14.234 となり，一般的な基準である 10 より大きいので，多重共線性の存在が示唆される。

重回帰分析による予測結果は図 4 のようになる。

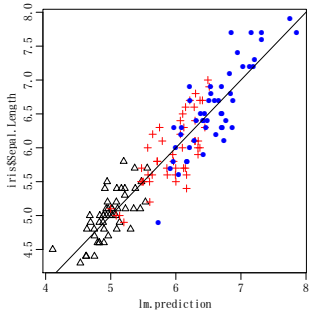

図 4　重回帰分析による予測結果

PCR 回帰の結果は以下のように示される。

‘1 comps’, ‘2 comps’, ‘3 comps’ はそれぞれ，「第 1 主成分だけ」，「第 2 主成分まで」，「第 3 主成分まで」を考慮した結果ということである。X の行の数字は 3 つの主成分の分散の累積寄与率，Sepal.Length 行の数値は，Sepal.Length の分散がそれぞれの予測値の分散で説明される割合（決定係数）である。

```
Data:         X dimension: 150 3
        Y dimension: 150 1
Fit method: svdpc
Number of components considered: 3
TRAINING: % variance explained
              1 comps  2 comps  3 comps
X               95.09    99.12   100.00
Sepal.Length    74.29    82.11    85.86
```

予測値の最初の部分を示す。

```
    1 comps  2 comps  3 comps
1  4.880974 4.987828 5.015416
2  4.899464 4.717888 4.689997
3  4.858013 4.789387 4.749251
4  4.929821 4.808353 4.825994
5  4.877276 5.041815 5.080499
6  4.996961 5.360115 5.377194
7  4.898979 4.957293 4.894684
8  4.918727 4.970316 5.021245
9  4.903162 4.663900 4.624913
10 4.915514 4.784900 4.881642
```


- 第 1 主成分だけを用いた分析結果（予測値）

第 1 主成分だけを用いて予測すると図 5 のようになる。

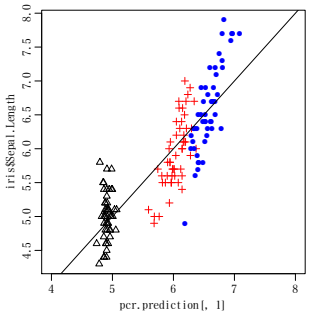

図 5　第 1 主成分だけを用いた予測値

- 第 1，第 2 主成分だけを用いた分析結果（予測値）

第 1，第 2 主成分だけを用いて予測すると図 6 のようになる。

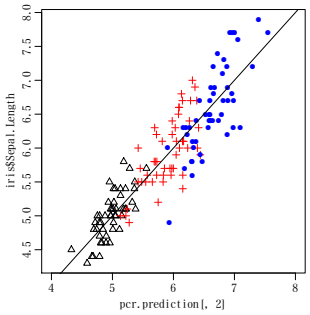

図 6　第 1，第 2 主成分だけを用いた予測値

- 第 1 〜 第 3 主成分（全部）を用いた分析結果（予測値）

第 1，第 2，第 3 主成分の全部を用いて予測すると図 7 のようになる。

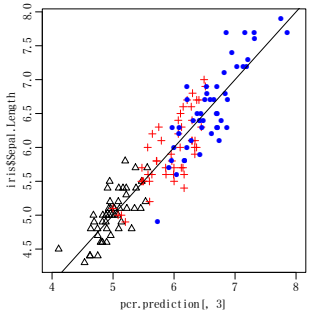

図 7　第 1 〜 第 3 主成分を用いた予測値

この分析結果，図 7（pcr.prediction[,3]）は，3 つの独立変数を用いた重回帰分析の結果，図 4（lm.prediction）と，全く同じである。

以上のように，分析方法が違っても，元々のデータが持つ情報量の全部を用いれば結果は同じということである（そうでなければ話がおかしくなる。理論的に元の情報全てを抽出できない分析法は生き残れない）。

pcr 関数を使わずに分析する

まず，重回帰分析に使った 3 つの独立変数の主成分分析を行い，それぞれの主成分得点を求める。

主成分の分散（固有値）は，それぞれ 3.696，0.157，0.034 で，第 1 主成分の寄与率は 95.1% と，圧倒的に大きい。

主成分得点の最初の方を示す。

```
           PC1    PC2    PC3
 [1,] -2.592 -0.183  0.032
 [2,] -2.543  0.311 -0.032
 [3,] -2.654  0.117 -0.046
 [4,] -2.461  0.208  0.020
 [5,] -2.602 -0.281  0.044
 [6,] -2.280 -0.621  0.020
 [7,] -2.544 -0.100 -0.072
 [8,] -2.491 -0.088  0.059
 [9,] -2.533  0.409 -0.045
[10,] -2.499  0.223  0.111
```


表 20 に第 1 主成分得点だけを使って，重回帰分析を行った結果を示す。

表 20　第 1 主成分得点だけを使って，重回帰分析

	偏回帰係数	標準誤差	t 値	P 値	標準化偏回帰係数	VIF
prcomp.predict[, 1]	0.371	0.018	20.680	< 0.001	0.862	1.000
定数項	5.843	0.034	169.873	< 0.001		

重相関係数 $R = 0.862$
重相関係数の二乗（決定係数）$R^2 = 0.743$
自由度調整済み重相関係数の二乗 = 0.741
回帰の分散分析：F 値 $(1, 148) = 427.642$，P 値 < 0.001

表 21 に第 1，第 2 主成分得点だけを使って，重回帰分析を行った結果を示す。
表の 1 行目は，表 20 の 1 行目と全く同じであることに注目。

表 21　第 1，第 2 主成分得点だけを使って，重回帰分析

	偏回帰係数	標準誤差	t 値	P 値	標準化偏回帰係数	VIF
prcomp.predict[, 1:2]PC1	0.371	0.015	24.704	< 0.001	0.862	1.000
prcomp.predict[, 1:2]PC2	−0.585	0.073	−8.014	< 0.001	−0.280	1.000
定数項	5.843	0.029	202.936	< 0.001		

重相関係数 $R = 0.906$
重相関係数の二乗（決定係数）$R^2 = 0.821$
自由度調整済み重相関係数の二乗 = 0.819
回帰の分散分析：F 値 $(2, 147) = 337.263$，P 値 < 0.001

表 22 に 第 1 〜 第 3 主成分得点を使って，重回帰分析を行った結果を示す。
表の 1，2 行目は，表 21 の 1，2 行目と全く同じであることに注目。

表 22　第 1 〜 第 3 主成分得点を使って，重回帰分析

	偏回帰係数	標準誤差	t 値	P 値	標準化偏回帰係数	VIF
prcomp.predict[, 1:3]PC1	0.371	0.013	27.697	< 0.001	0.862	1.000
prcomp.predict[, 1:3]PC2	−0.585	0.065	−8.984	< 0.001	−0.280	1.000
prcomp.predict[, 1:3]PC3	0.870	0.140	6.227	< 0.001	0.194	1.000
定数項	5.843	0.026	227.519	< 0.001		

重相関係数 $R = 0.927$
重相関係数の二乗（決定係数）$R^2 = 0.859$
自由度調整済み重相関係数の二乗 = 0.856
回帰の分散分析：F 値 $(3, 146) = 295.539$，P 値 < 0.001

PCR 回帰の結果と主成分得点を使って重回帰分析をした結果の比較（先頭の 10 ケース）

表 23 の，左の 3 列は主成分得点（第 1 〜 3 主成分得点）を使った重回帰分析の予測値，右の 3 列は pcr 回帰（第 1 〜 3 主成分）を使った予測値である。それぞれ対応する予測値が同じであることがわかる。

つまり，主成分回帰と，主成分得点を用いる重回帰分析は同じだということである。

表 23　PCR 回帰の結果と主成分得点を使って重回帰分析をした結果の比較

	V1	V2	V3	1 comps	2 comps	3 comps
1	4.88	4.99	5.02	4.88	4.99	5.02
2	4.90	4.72	4.69	4.90	4.72	4.69
3	4.86	4.79	4.75	4.86	4.79	4.75
4	4.93	4.81	4.83	4.93	4.81	4.83
5	4.88	5.04	5.08	4.88	5.04	5.08
6	5.00	5.36	5.38	5.00	5.36	5.38
7	4.90	4.96	4.89	4.90	4.96	4.89
8	4.92	4.97	5.02	4.92	4.97	5.02
9	4.90	4.66	4.62	4.90	4.66	4.62
10	4.92	4.78	4.88	4.92	4.78	4.88

実は，この現象を正しく説明するには，偏相関係数を考えればよい。偏相関係数は相関関係を見ようとしている二変数以外の変数の影響を取り除いた，「正味」の相関を表す。具体的には，以下のように，x_1，x_2，x_4 を用いて，それぞれ y，x_3 を予測しその残差を求め，残差同士の相関係数を求めると，それが y と x_3 の偏相関係数である（実際には，このようにして馬鹿正直に求める必要はない）。

```
> y.adj <- lm(y ~ x1+x2+x4, d)$residuals
> x3.adj <- lm(x3 ~ x1+x2+x4, d)$residuals
> cor(y.adj, x3.adj)
[1] -0.2028805
```


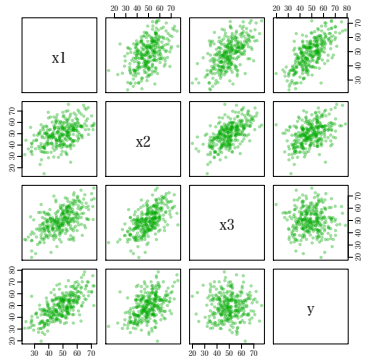

図 8　データの全容

1.4　多変数を考慮しなければならない例

1.4.1　抑制変数

抑制変数とは，従属変数との単相関係数の符号と重回帰分析の偏回帰係数の符号が逆になる独立変数のことである。

独立変数が x_1 〜 x_4 の 4 個，従属変数が y というデータがある。変数間の相関係数は，表 24 のようになっていた。

表 24　単相関係数

	x_1	x_2	x_3	x_4	y
x_1	1.000	0.459	0.375	0.223	0.464
x_2	0.459	1.000	0.585	0.220	0.544
x_3	0.375	0.585	1.000	0.205	0.243
x_4	0.223	0.220	0.205	1.000	0.426
y	0.464	0.544	0.243	0.426	1.000

このデータにおいて，x_1 〜 x_4 を用いて y を予測する重回帰式を求める。

表 26 の下三角行列に偏相関係数を示す（上三角行列は単相関係数，対角成分は重相関係数）。独立変数 x_3 と従属変数 y の偏相関係数は −0.203 で，負の値になっている。

表 26　単相関係数と偏相関係数

	x_1	x_2	x_3	x_4	y
x_1	0.549	0.459	0.375	0.223	0.464
x_2	0.137	0.726	0.585	0.220	0.544
x_3	0.185	0.522	0.622	0.205	0.243
x_4	0.010	−0.098	0.146	0.447	0.426
y	0.284	0.442	−0.203	0.370	0.676

従属変数 y とそこそこの相関を示す x_1 と x_2 を用いると，表 28 のような結果が得られる。自由度調整済みの重相関係数の二乗も 0.481 であり，大して大きな値ではない。

表 28　相関がきわめて低い独立変数を含まない重回帰モデル model1

	偏回帰係数	標準誤差	t 値	P 値	標準化偏回帰係数	VIF
x1	0.630	0.048	13.002	< 0.001	0.610	1.271
x2	0.162	0.049	3.321	0.001	0.156	1.271
定数項	10.351	2.527	4.096	< 0.001		

重相関係数 $R = 0.696$
重相関係数の二乗（決定係数）$R^2 = 0.485$
自由度調整済み重相関係数の二乗 = 0.481
回帰の分散分析：F 値 $(2, 297) = 139.797$，P 値 < 0.001

表 25　抑制変数が含まれる重回帰モデル

	偏回帰係数	標準誤差	t 値	P 値	標準化偏回帰係数	VIF
x1	0.250	0.061	4.137	< 0.001	0.250	1.317
x2	0.473	0.069	6.875	< 0.001	0.473	1.703
x3	−0.191	0.066	−2.893	0.004	−0.191	1.564
x4	0.305	0.055	5.569	< 0.001	0.305	1.079
定数項	8.095	3.794	2.134	0.034		

重相関係数 $R = 0.676$
重相関係数の二乗（決定係数）$R^2 = 0.457$
自由度調整済み重相関係数の二乗 = 0.446
回帰の分散分析：F 値 $(4, 195) = 41.068$，P 値 < 0.001

結果は表 25 のようになる。独立変数 x_3 と従属変数 y の単相関係数は 0.243 と正の値であるにもかかわらず，偏回帰係数は −0.191 となり負の値になっている。このような状況になると，多くの人は「多重共線性のせいだろう」と考えるかもしれないが，VIF を見てわかるように，多重共線性はないようである。

1.4.2　無相関の独立変数を分析に含める意味

前節の抑制変数の場合に含まれることではあるが，従属変数と相関関係のない独立変数を重回帰モデルに含めることについて考える。普通は誰しも「相関のない変数など含める必要はない」と思うであろうが，必ずしも適切ではない。

図 8 および表 27 に示すようなデータでは，y と x_3 の相関係数は 0.018 なので，重回帰モデルに含む必要はないように思われる。

表 27　相関係数行列

	x_1	x_2	x_3	y
x_1	1.000	0.462	0.559	0.682
x_2	0.462	1.000	0.589	0.438
x_3	0.559	0.589	1.000	0.018
y	0.682	0.438	0.018	1.000

y との相関が低い x_3 をモデルに取り入れると，表 29 のように自由度調整済みの重相関係数の二乗は 0.798 となり，大幅に大きくなる。y と x_3 の相関係数は 0.018 だが，偏相関係数は実は −0.782 もあったのだ。

表 29　相関がきわめて低い独立変数を含む重回帰モデル model2

	偏回帰係数	標準誤差	t 値	P 値	標準化偏回帰係数	VIF
x1	0.915	0.033	27.714	< 0.001	0.887	1.514
x2	0.494	0.034	14.451	< 0.001	0.475	1.593
x3	−0.793	0.037	−21.571	< 0.001	−0.758	1.824
定数項	19.512	1.635	11.936	< 0.001		

重相関係数 $R = 0.894$
重相関係数の二乗（決定係数）$R^2 = 0.800$
自由度調整済み重相関係数の二乗 = 0.798
回帰の分散分析：F 値 $(3, 296) = 393.995$，P 値 < 0.001

実測値と予測値の対応は図 9 のようになる。

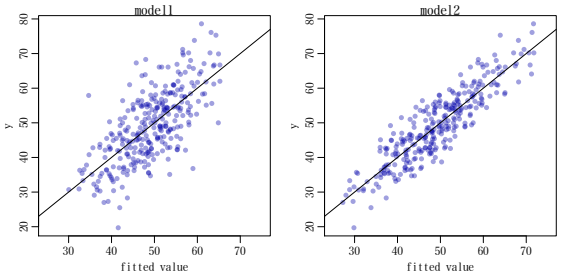

図 9　2 つのモデルによる予測値と実測値の比較

1.4.3　判別分析において二群で平均値が同じ変数は使うか？

判別分析の場合に，各群で平均値が（ほとんど）同じになる変数は説明変数として役に立つだろうか？

実際にそのようなデータがあるかどうかは別として，以下のような人工的なデータを考えれば，たとえ平均値が同じ変数でも，説明変数に加えた方がよい場合もあることが分かる。

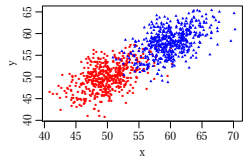

図 10　平均値に差のない変数を除いた場合の散布図

青で示したデータは，赤で示したデータ x，y，z の 3 変数をそれぞれ，$(+10, +8, +0)$ 平行移動したものである。

x と y だけを使って分類しようとしても，中央付近で青と赤は混じり合っていることがわかる（正判別率は 0.946 である）。

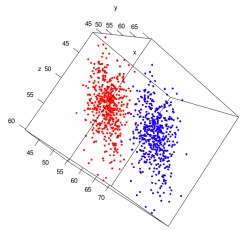

図 11　平均値に差のない変数も含んだ場合の三次元散布図

平均値に差のない z も考慮に入れて三次元散布図を描くと，ある平面で青と赤はかなり明確に判別できることがわかる（正判別率は 0.995 である）。

1.5　ロジスティック回帰分析における変数選択

医学系の論文で，ロジスティック回帰分析を用いて分析する場合に，まず単変量解析を行って，有意であった変数を寄せ集めて最終的なロジスティック回帰分析を行うというのがお手本のようになっているようだ。

重回帰分析の場合にも同じような過程をたどる人はそれに比べると少ないような気がする。

1.5.1　重回帰分析における変数選択

重回帰分析において変数選択を行うオプションを持つ統計解析プログラムは少なくない。変数選択をしようとする人は，最初からそのオプションを採用するだろう。

例えば，10 個の独立変数候補から AIC による変数選択は R では以下のようになる*2。

```
Start:  AIC=-244.04
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10

       Df Sum of Sq     RSS     AIC
- x1    1    0.0001  6.9923 -246.04
- x2    1    0.0005  6.9928 -246.03
- x7    1    0.1136  7.1059 -244.42
- x9    1    0.1157  7.1080 -244.40
<none>               6.9923 -244.04
- x10   1    0.1828  7.1751 -243.46
```

*2 偏回帰係数の P 値に基づくステップワイズ変数選択法は，http://aoki2.si.gunma-u.ac.jp/R/sreg.html を参照。


```
- x5    1    0.2013  7.1936 -243.20
- x8    1    0.3506  7.3429 -241.14
- x3    1    0.8189  7.8112 -234.96
- x6    1    1.4884  8.4807 -226.74
- x4    1    3.1948 10.1870 -208.41

Step:  AIC=-246.04
y ~ x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10

       Df Sum of Sq     RSS     AIC
- x2    1    0.0005  6.9928 -248.03
- x7    1    0.1150  7.1073 -246.41
- x9    1    0.1158  7.1081 -246.39
<none>               6.9923 -246.04
- x10   1    0.1857  7.1780 -245.41
- x5    1    0.2051  7.1974 -245.15
- x8    1    0.3508  7.3431 -243.14
- x3    1    0.8589  7.8513 -236.45
- x6    1    2.4964  9.4888 -217.51
- x4    1    3.2020 10.1943 -210.33

Step:  AIC=-248.03
y ~ x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10

       Df Sum of Sq     RSS     AIC
- x9    1    0.1162  7.1090 -248.38
<none>               6.9928 -248.03
```



```
- x7    1    0.1599  7.1526 -247.77
- x10   1    0.1854  7.1782 -247.41
- x5    1    0.2107  7.2035 -247.06
- x8    1    0.3893  7.3821 -244.61
- x3    1    0.9346  7.9274 -237.48
- x4    1    3.2264 10.2192 -212.09
- x6    1    3.3243 10.3171 -211.14

Step:  AIC=-248.38
y ~ x3 + x4 + x5 + x6 + x7 + x8 + x10

       Df Sum of Sq     RSS     AIC
<none>               7.1090 -248.38
- x7    1    0.1506  7.2596 -248.28
- x10   1    0.1793  7.2884 -247.89
- x5    1    0.2011  7.3101 -247.59
- x8    1    0.3225  7.4315 -245.94
- x3    1    1.2041  8.3132 -234.73
- x6    1    3.2132 10.3222 -213.09
- x4    1    3.2554 10.3644 -212.68

Call:
lm(formula = y ~ x3 + x4 + x5 + x6 + x7 + x8 + x10, data = d)

Coefficients:
(Intercept)           x3           x4           x5           x6           x7
  8.621e-17    4.542e-01   -2.211e-01    8.413e-02    4.419e-01    8.712e-02
```



```
          x8           x10
-1.138e-01    7.544e-02
```

最終的なモデルは，表 30 に示すものになる。

表 30　AIC による変数選択により選ばれた変数

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0000	0.0278	0.00	1.0000
x3	0.4542	0.1151	3.95	0.0002
x4	−0.2211	0.0341	−6.49	0.0000
x5	0.0841	0.0522	1.61	0.1101
x6	0.4419	0.0685	6.45	0.0000
x7	0.0871	0.0624	1.40	0.1661
x8	−0.1138	0.0557	−2.04	0.0439
x10	0.0754	0.0495	1.52	0.1311

x_5, x_7, x_{10} の偏回帰係数は 5% の有意水準のもとでは有意なものではないがモデルには含まれることになった。

もし，医学論文における分析手順を踏襲するとすれば，まずは以下のように単回帰分析を繰り返すことになる。

表 31　x_1 を使った単回帰分析

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0000	0.0689	0.00	1.0000
x1	0.7280	0.0693	10.51	0.0000

表 32　x_2 を使った単回帰分析

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0000	0.0998	0.00	1.0000
x2	0.1189	0.1003	1.18	0.2389

これらの結果をまとめると，表 33 のようになる。つまり，x_2，x_4 以外の独立変数は，有意な偏回帰係数ということになる。

表 33　単回帰分析結果のまとめ

	Estimate	Std. Error	t value	Pr(>|t|)
x1	0.7280	0.0693	10.51	0.0000
x2	0.1189	0.1003	1.18	0.2389
x3	0.8874	0.0466	19.05	0.0000
x4	−0.0249	0.1010	−0.25	0.8058
x5	0.7779	0.0635	12.25	0.0000
x6	0.9239	0.0387	23.90	0.0000
x7	0.8127	0.0589	13.81	0.0000
x8	0.7137	0.0708	10.09	0.0000
x9	0.5003	0.0875	5.72	0.0000
x10	0.7464	0.0672	11.10	0.0000

なお，独立変数の偏回帰係数の検定は独立変数と従属変数の相関係数の検定と同じである。単回帰解析を繰り返す必要はなく，相関係数の検定を行うだけで十分なわけである。

表 34　相関係数行列

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
x1	1.000	0.289	0.613	−0.089	0.491	0.814	0.561	0.508	0.267	0.493	0.728
x2	0.289	1.000	−0.013	−0.172	0.114	0.230	−0.200	0.018	−0.105	−0.041	0.119
x3	0.613	−0.013	1.000	0.308	0.831	0.888	0.869	0.853	0.614	0.800	0.887
x4	−0.089	−0.172	0.308	1.000	0.254	0.088	0.148	0.286	0.265	0.208	−0.025
x5	0.491	0.114	0.831	0.254	1.000	0.754	0.696	0.672	0.478	0.731	0.778
x6	0.814	0.230	0.888	0.088	0.754	1.000	0.767	0.737	0.453	0.685	0.924
x7	0.561	−0.200	0.869	0.148	0.696	0.767	1.000	0.799	0.540	0.761	0.813
x8	0.508	0.018	0.853	0.286	0.672	0.737	0.799	1.000	0.612	0.679	0.714
x9	0.267	−0.105	0.614	0.265	0.478	0.453	0.540	0.612	1.000	0.485	0.500
x10	0.493	−0.041	0.800	0.208	0.731	0.685	0.761	0.679	0.485	1.000	0.746
y	0.728	0.119	0.887	−0.025	0.778	0.924	0.813	0.714	0.500	0.746	1.000

従属変数との相関係数の P 値が 0.05 以下である変数だけをピックアップした重回帰モデル（表 35）では，偏回帰係数が有意な独立変数は x_6，x_7，x_8 だけということになり，先に示した AIC に基づく変数選択を行ったモデル（表 30）とは異なる。

表 35　単回帰分析で有意な変数のみを寄せ集めても適切な重回帰式は作れない

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0000	0.0335	0.00	1.0000
x1	0.0183	0.0650	0.28	0.7785
x3	0.0976	0.1359	0.72	0.4748
x5	0.0825	0.0635	1.30	0.1975
x6	0.6358	0.1140	5.58	0.0000
x7	0.2004	0.0726	2.76	0.0070
x8	−0.1575	0.0682	−2.31	0.0232
x9	0.0511	0.0450	1.13	0.2598
x10	0.0930	0.0600	1.55	0.1244

このような不合理な変数選択が行われている理由の 1 つは，サンプルサイズである。サンプルサイズを増やしていけば，従属変数と独立変数の相関係数がほとんど同じであっても相関係数（偏回帰係数）は「統計学的に有意」となるからである。

サンプルサイズが大きいと，ほとんどの独立変数は従属変数と「有意な相関」を持つから重回帰モデルにはほとんど全ての独立変数を含めようという判断になるが，その判断が正しいという保証はほとんどない。

独立変数候補の全てを対象として変数選択をするのが正しい手順である。

1.5.2　ロジスティック回帰分析における変数選択

前節のデータでの従属変数は連続変数であった。これをある基準で 0/1 データに切り分けて分析を行うのがロジスティック回帰分析であるから，変数選択も「独立変数候補を全て対象として変数選択をする」のが正しい手順となる。

しかし，ロジスティック回帰分析において変数選択を行える統計解析プログラムがない，あるいは使用している統計解析プログラムがそのような機能を備えていないというような場合には，前節の前半に書いたような「間違った手順」を採用するしかないということになるのであろう。

分析に使用するデータは，前節のデータセットの y を二値変数にカテゴリー化したものである。相関係数行列を表 36 に示す。

表 36　ロジスティック回帰分析に使用される変数間の相関係数行列

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
x1	1.000	0.289	0.613	−0.089	0.491	0.814	0.561	0.508	0.267	0.493	0.551
x2	0.289	1.000	−0.013	−0.172	0.114	0.230	−0.200	0.018	−0.105	−0.041	0.061
x3	0.613	−0.013	1.000	0.308	0.831	0.888	0.869	0.853	0.614	0.800	0.698
x4	−0.089	−0.172	0.308	1.000	0.254	0.088	0.148	0.286	0.265	0.208	−0.026
x5	0.491	0.114	0.831	0.254	1.000	0.754	0.696	0.672	0.478	0.731	0.667
x6	0.814	0.230	0.888	0.088	0.754	1.000	0.767	0.737	0.453	0.685	0.725
x7	0.561	−0.200	0.869	0.148	0.696	0.767	1.000	0.799	0.540	0.761	0.617
x8	0.508	0.018	0.853	0.286	0.672	0.737	0.799	1.000	0.612	0.679	0.564
x9	0.267	−0.105	0.614	0.265	0.478	0.453	0.540	0.612	1.000	0.485	0.360
x10	0.493	−0.041	0.800	0.208	0.731	0.685	0.761	0.679	0.485	1.000	0.554
y	0.551	0.061	0.698	−0.026	0.667	0.725	0.617	0.564	0.360	0.554	1.000

単変量ロジスティック回帰分析の結果をまとめると表 37 のようになる。

表 37　単変量ロジスティック回帰分析結果のまとめ

	Estimate	Std. Error	z value	Pr(>|z|)
x1	1.961	0.427	4.590	0.000
x2	0.135	0.222	0.609	0.542
x3	3.549	0.747	4.748	0.000
x4	−0.058	0.224	−0.259	0.795
x5	2.505	0.528	4.744	0.000
x6	5.950	1.514	3.931	0.000
x7	2.334	0.478	4.879	0.000
x8	1.931	0.430	4.490	0.000
x9	0.866	0.258	3.361	0.001
x10	2.143	0.484	4.423	0.000

P 値が 0.05 以下である変数だけをピックアップした多変量ロジスティック回帰モデルは，x_2，x_4 以外を含むものである（表 38）。しかし，このモデルで有意な偏回帰係数を持つ独立変数は x_6 だけということになってしまう。

表 38　単変量回帰で変数選択した多変量ロジスティックモデル

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−4.4865	1.6407	−2.73	0.0062
x1	0.4929	1.1956	0.41	0.6801
x3	0.7566	2.1847	0.35	0.7291
x5	2.0050	1.2284	1.63	0.1026
x6	6.0954	2.8329	2.15	0.0314
x7	1.6277	1.4889	1.09	0.2743
x8	−0.6226	1.4308	−0.44	0.6635
x9	0.0036	0.7745	0.00	0.9963
x10	−0.0520	1.2943	−0.04	0.9679

正しい分析手順は，全ての独立変数を候補とするのが出発点である。しかし，このデータセットではいくつかの独立変数をモデルに含めると数値計算上の問題が生じるので，それらを除いたモデルをスタート地点とする。

```
Start:  AIC=42.48
y ~ x1 + x2 + x3 + x4 + x5 + x7 + x8 + x9

       Df Deviance    AIC
- x2    1   24.605 40.605
- x8    1   24.696 40.696
- x7    1   24.917 40.917
- x9    1   25.541 41.541
<none>      24.483 42.483
- x5    1   27.176 43.176
- x1    1   29.428 45.428
- x4    1   32.988 48.988
- x3    1   36.955 52.955

Step:  AIC=40.61
y ~ x1 + x3 + x4 + x5 + x7 + x8 + x9

       Df Deviance    AIC
- x8    1   25.009 39.009
- x7    1   25.414 39.414
- x9    1   25.954 39.954
<none>      24.605 40.605
- x5    1   27.357 41.357
```



```
- x1    1   30.077 44.077
- x4    1   33.780 47.780
- x3    1   37.158 51.158

Step:  AIC=39.01
y ~ x1 + x3 + x4 + x5 + x7 + x9

       Df Deviance    AIC
- x7    1   25.446 37.446
- x9    1   25.976 37.976
<none>      25.009 39.009
- x5    1   27.433 39.433
- x1    1   30.103 42.103
- x4    1   34.462 46.462
- x3    1   37.455 49.455

Step:  AIC=37.45
y ~ x1 + x3 + x4 + x5 + x9

       Df Deviance    AIC
- x9    1   26.419 36.419
<none>      25.446 37.446
- x5    1   28.046 38.046
- x1    1   31.179 41.179
- x4    1   34.521 44.521
- x3    1   42.187 52.187
```



```
Step:  AIC=36.42
y ~ x1 + x3 + x4 + x5

       Df Deviance    AIC
<none>      26.419 36.419
- x5    1   29.143 37.143
- x1    1   33.416 41.416
- x4    1   35.324 43.324
- x3    1   43.869 51.869

Call:  glm(formula = y ~ x1 + x3 + x4 + x5, family = binomial, data = d)

Coefficients:
(Intercept)           x1           x3           x4           x5
     -3.770        2.240        4.984       -2.164        1.473

Degrees of Freedom: 99 Total (i.e. Null);  95 Residual
Null Deviance:            118.6
Residual Deviance: 26.42         AIC: 36.42
```

最終的なモデルは，表 39 のようになった。x_5 は有意な偏回帰係数を持たないが，モデルに入っている。

表 39　最終的な多変量ロジスティックモデル

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−3.7699	1.1413	−3.30	0.0010
x1	2.2400	1.0131	2.21	0.0270
x3	4.9840	1.7332	2.88	0.0040
x4	−2.1640	0.9295	−2.33	0.0199
x5	1.4733	0.9632	1.53	0.1261

なお，もう一つの変数選択法である「総当たり法[*3]」によって求めた結果は，表 40 のようになる。

表 40　AIC が上位 10 位に入るロジスティック回帰分析の結果（総当たり法）

AIC	Formula
36.42	y ~ x1 + x3 + x4 + x5
37.14	y ~ x1 + x3 + x4
37.45	y ~ x1 + x3 + x4 + x5 + x9
37.76	y ~ x1 + x2 + x3 + x4 + x5
37.98	y ~ x1 + x3 + x4 + x5 + x7
38.05	y ~ x1 + x3 + x4 + x9
38.38	y ~ x1 + x3 + x4 + x5 + x8
38.58	y ~ x1 + x2 + x3 + x4
38.64	y ~ x1 + x3 + x4 + x7
38.89	y ~ x1 + x3 + x4 + x8

[*3] http://aoki2.si.gunma-u.ac.jp/R/all.logistic.html

目次

1　重回帰分析において 2

1.1　重回帰分析における誤解 2

1.1.1　「順序尺度データは重回帰分析には使えない」というのは誤り 2

1.1.2　「回帰の分散分析の P 値が有意ならば優れたモデルである」というのは誤り 6

1.1.3　「独立変数間の相関係数に 0.8 を超えるものがあると多重共線性が生じる」というのは誤り . 9

1.2　多重共線性の回避策 13

1.3　主成分回帰 23

1.3.1　主成分回帰と直線回帰の違い 24

1.3.2　独立変数の中に相関の高いものがある場合の実例 25

1.4　多変数を考慮しなければならない例 38

1.4.1　抑制変数 38

1.4.2　無相関の独立変数を分析に含める意味 42

1.4.3　判別分析において二群で平均値が同じ変数は使うか？ 47

1.5　ロジスティック回帰分析における変数選択 49

1.5.1　重回帰分析における変数選択 49

1.5.2　ロジスティック回帰分析における変数選択 57