第二十章　重回帰分析（R で）

重回帰分析では、2 つ以上の独立変数 X で従属変数 Y を説明できるかどうか評価する。回帰式を拡張した $Y=\alpha+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+,,,+e$ という線形モデルを考える。その際、e(残差)の二乗和が最小になるような α、β_i を推定するのが一般的である。

例題 20.1　以下のような 5 種類の変数(X_1〜X_4, Y)の調査データ（n = 33）がある（33×5 の行列データ[i]）。講義用の Web ページからダウンロードしておく[ii]。

	変数(Xi)				
j	X_1(℃)	X_2(cm)	X_3(mm)	X_4(min)	Y(ml)
1	6	9.9	5.7	1.6	2.12
2	1	9.3	6.4	3	3.39
3	-2	9.4	5.7	3.4	3.61
4	11	9.1	6.1	3.4	1.72
5	-1	6.9	6	3	1.8
6	2	9.3	5.7	4.4	3.21
7	5	7.9	5.9	2.2	2.59
8	1	7.4	6.2	2.2	3.25
9	1	7.3	5.5	1.9	2.86
10	3	8.8	5.2	0.2	2.32
11	11	9.8	5.7	4.2	1.57
12	9	10.5	6.1	2.4	1.5
13	5	9.1	6.4	3.4	2.69
14	-3	10.1	5.5	3	4.06
15	1	7.2	5.5	0.2	1.98
16	8	11.7	6	3.9	2.29
17	-2	8.7	5.5	2.2	3.55
18	3	7.6	6.2	4.4	3.31
19	6	8.6	5.9	0.2	1.83
20	10	10.9	5.6	2.4	1.69
21	4	7.6	5.8	2.4	2.42
22	5	7.3	5.8	4.4	2.98
23	5	9.2	5.2	1.6	1.84
24	3	7	6	1.9	2.48
25	8	7.2	5.5	1.6	2.83
26	8	7	6.4	4.1	2.41
27	6	8.8	6.2	1.9	1.78
28	6	10.1	5.4	2.2	2.22
29	3	12.1	5.4	4.1	2.72
30	5	7.7	6.2	1.6	2.36
31	1	7.8	6.8	2.4	2.81
32	8	11.5	6.2	1.9	1.64
33	10	10.4	6.4	2.2	1.82

以下で、R を操作するコードは赤字（ファイル名や変数名などユーザーが変更する箇所は緑で網掛け）で、R の出力結果は青字で示す。

```
> d <- read.table("Ex20.1.txt", header=T)     #例データの読み込み
> cor(d)                                       #相関係数行列の確認（必須ではない）
```

	X1	X2	X3	X4	Y
X1	1.00000000	0.3287174	0.16767376	0.05191106	-0.73080887
X2	0.32871738	1.0000000	-0.14549623	0.18032591	-0.21203749
X3	0.16767376	-0.1454962	1.00000000	0.24133905	-0.05540952
X4	0.05191106	0.1803259	0.24133905	1.00000000	0.31266707
Y	-0.73080887	-0.2120375	-0.05540952	0.31266707	1.00000000

例データの 5 つの変数間の相関係数で、X1 と Y との間に-0.73 の関係がみえるが、それ以外の変数間の相関係数は大きくないことがわかる。相関係数行列は、重回帰分析に必須ではないが、従属変数を説明する独立変数間に強い相関がみられる場合に注意する。

lm 関数を用いて重回帰分析を行う。$Y=\alpha+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4$ という重回帰モデル（モデル 16 としておく）に対し、帰無仮説（H_0）：$\beta_1=\beta_2=\beta_3=\beta_4=0$ を検定する[iii]。R コードは以下のようになる。

```
> model16 <- lm(Y ~ X1 + X2 + X3 + X4, data =d)
> summary(model16)
```

lm 関数の左端に従属変数を置き、”~”記号の右辺に従属変数を並べている。この 4 変数を独立変数とする重回帰モデルを model16 とし、各独立変数 X_i の係数(β_i)を推定し、すべての係数が 0 になる（独立変数は従属変数を説明しない）か否かを検定する。

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4, data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-1.5070   -0.1112    0.0401    0.1550    0.9617

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  2.95828    1.36361    2.169    0.03869 *
X1          -0.12932    0.02129   -6.075    1.5e-06 ***
X2          -0.01878    0.05628   -0.334    0.74102
X3          -0.04621    0.20727   -0.223    0.82518
X4           0.20876    0.06703    3.114    0.00423 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4238 on 28 degrees of freedom Multiple R-squared: 0.6589, Adjusted R-squared:
0.6102   F-statistic: 13.52 on 4 and 28 DF,   p-value: 2.948e-06
```

結果より、重回帰モデルは $Y=2.95828-0.12932X_1-0.01878X_2-0.04621X_3+0.20876X_4$(式 1)となった。下段にある Adjusted R-squared（自由度調整済み R^2 値）：0.6102 はこの重回帰モデルの当てはまりの良さを示す数値で、式 1 によりデータの 61.02%が説明できることを示す。また、推定された係数のうち Y 切片は 5%水準で有意となり、X1 と X4 の回帰係数はそれぞ

れ 0.1%水準と 1%水準で有意となった。p-value（p 値）は帰無仮説が有意水準 2.948×10^{-6} で棄却されることを意味する。ただし、重回帰分析のような多変量解析では、統計的検定自体はあまり意味がない。重回帰分析の場合、適切なモデルの選択（従属変数を説明する独立変数の選択）が重要である。例データの場合、最大 4 つの独立変数があるので、各変数が従属変数の説明に有効であるか否かに着目すると 2^4 通りの重回帰モデルが考えられる[iv]。

$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ … 変数 $X_1 \sim X_4$
$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ … 変数 $X_1 \sim X_3$
$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4$ … 変数 X_1, X_2, X_4
$Y = \alpha + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4$ … 変数 X_1, X_3, X_4
$Y = \alpha + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$ … 変数 $X_2 \sim X_4$
：　　：
$Y = \alpha$

それぞれのモデルを解析し、適したモデルを選択すればよい。モデル選択に便利な指標に AIC[v]があり、R で step 関数を使えば、AIC に基づくモデル選択を行える。以下では MuMIn パッケージ[vi]を使ったモデル選択法を示す。

- library(MuMIn)
- options(na.action = "na.fail")　#オプション設定
- dredge(model16, rank="AIC")　#上述の全 4 種の変量を組み込んだ回帰モデルを指定

```
Fixed term is "(Intercept)"
Global model call: lm(formula = Y ~ X1 + X2 + X3 + X4, data = d)
---
Model selection table
   (Intrc)      X1       X2       X3     X4 df  logLik  AIC delta weight
10   2.552 -0.1324                   0.2013  4 -15.863 39.7  0.00  0.516
12   2.673 -0.1305 -0.01542           0.2045  5 -15.816 41.6  1.90  0.199
14   2.707 -0.1319          -0.02768 0.2035  5 -15.852 41.7  1.98  0.192
16   2.958 -0.1293 -0.01878 -0.04621 0.2088  6 -15.787 43.6  3.85  0.075
2    3.050 -0.1292                           3 -20.933 47.9  8.14  0.009
6    2.355 -0.1312            0.11970        4 -20.768 49.5  9.81  0.004
4    2.929 -0.1310  0.01444                  4 -20.902 49.8 10.08  0.003
8    2.064 -0.1344  0.02259  0.13740         5 -20.694 51.4 11.66  0.002
11   3.073         -0.12680          0.2077  4 -30.416 68.8 29.11  0.000
15   5.217         -0.14490 -0.35110 0.2398  5 -29.659 69.3 29.59  0.000
9    2.019                           0.1791  3 -31.838 69.7 29.95  0.000
13   3.387                  -0.24090 0.1983  4 -31.502 71.0 31.28  0.000
1    2.474                                   2 -33.535 71.1 31.34  0.000
3    3.335         -0.09689                  3 -32.776 71.6 31.83  0.000
5    3.039                  -0.09604         3 -33.484 73.0 33.24  0.000
7    4.286         -0.10280 -0.15270         4 -32.644 73.3 33.56  0.000
Models ranked by AIC(x)
```

AIC の小さい順に 16 種の回帰モデルが並んで出力される。AIC は、”より少ない独立変数で従属変数をうまく説明できる”モデルを探すための基準で、AIC が小さいほど効率的なモデルであると考える。一般に、AIC の差（delta 列）が 2 未満までのモデルはどれも優れたモデルとみなせ、例の場合は model10、model12、model14 が相当する、model12 は X2 を、model14 は X3 を model10 に加えたものなので、総合的に考えると model10 が最適モデルといえそうである。また、model1 は Y 切片のみの独立変数無しのモデルであり、model1 の AIC よりも最適モデルの AIC が充分低ければ、重回帰モデルによる従属変数の説明は意味があるといえる。

---

[i] Zar JH. 『Biostatistical Analysis.(5 版)』 (2009) Prentice Hall の p421

[ii] http://www.eonet.ne.jp/~vor-dem-gesetz/Ex20.1.txt

[iii] 対立仮説（H_1）は一つ以上の $\beta_i \neq 0$ （i=1～4）になる。

[iv] 各変数のが 1 次効果のみ考慮し（X^2,X^3,,の項を考えない）、交互作用は無視する

[v] 坂元ら『情報量基準』(1986) 共立出版

[vi] パッケージのインストール法は http://www.eonet.ne.jp/~vor-dem-gesetz/Rintro.pdf を参照