## 回帰分析の実践

### 重回帰分析

#### 基本形

「R 言語の基本」では，操作練習の一環として，身長データ data0.csv を用いて単回帰分析を行なってみたが，ここでは，より実践的な手法として，重回帰分析を実行してみる。データを読み取ったら，最初のデータを確認しておく。

```
> data0 = read.csv("data0.csv",header=TRUE)
> head(data0)
  No Height Sex Father Mother
1  1    162   1    161    147
2  2    165   1    156    152
3  3    167   1    168    158
4  4    167   1    167    158
5  5    168   1    166    158
6  6    168   1    160    155
```

まず，数式の書き方を拡張して，基本的な重回帰分析を実行してみよう。

```
> result1 = lm(Height ~ Father+Mother, data=data0)
> summary(result1)

Call:
lm(formula = Height ~ Father + Mother, data = data0)

Residuals:
     Min       1Q   Median       3Q      Max
-18.2351  -5.4696  -0.0506   4.6818  15.1843

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 68.09629   44.02308   1.547   0.1286
Father      -0.04101    0.20177  -0.203   0.8398
Mother       0.66977    0.25814   2.595   0.0126 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.147 on 47 degrees of freedom
Multiple R-squared:  0.1319,    Adjusted R-squared:  0.09501
F-statistic: 3.572 on 2 and 47 DF,  p-value: 0.03596
```

出力の見方は単回帰とほぼ同様で，説明変数に対応する回帰係数の数が増えている点だけが異なる。結果は，Father の（t 値に対する）p 値が高く，Height に与える Father の影響は「有意ではない」（Father の回帰係数は 0 とみなすのが妥当である）。一方，Mother の影響は「有意」である。しかしながら，説明力を意味する決定係数は非常に低く（約 10%），

このモデルでは十分ではない[1]，と全体的には評価できる。

**男女差を意識した分析**

ここで，未使用の Sex を活かすことを考える。そもそも，身長には男女差があることは，経験的には知られているので，散布図の書き方を工夫して，その様子を観察してみる。散布図のマーカーを性別に変えるために，plot の呼び出しに pch(マーカーの種類を決める)および col（描画色）の引数を追加し，それぞれ性別に異なる値が渡されるように，ifelse でコントロールしてみる。散布図行列で表示するために，数式は特定の縦軸を指定しない書き方 ~Height+Father+Mother を使用している。2 行目，3 行目の冒頭の+はプロンプトなので，入力する必要はない。

```
> plot( ~Height+Father+Mother, data=data0,
+   pch=ifelse(Sex==1,15,17),
+   col=ifelse(Sex==1,"blue","red"))
```

この結果，図 1 のような散布図が表示され，明らかに男女差があることが確認できる。

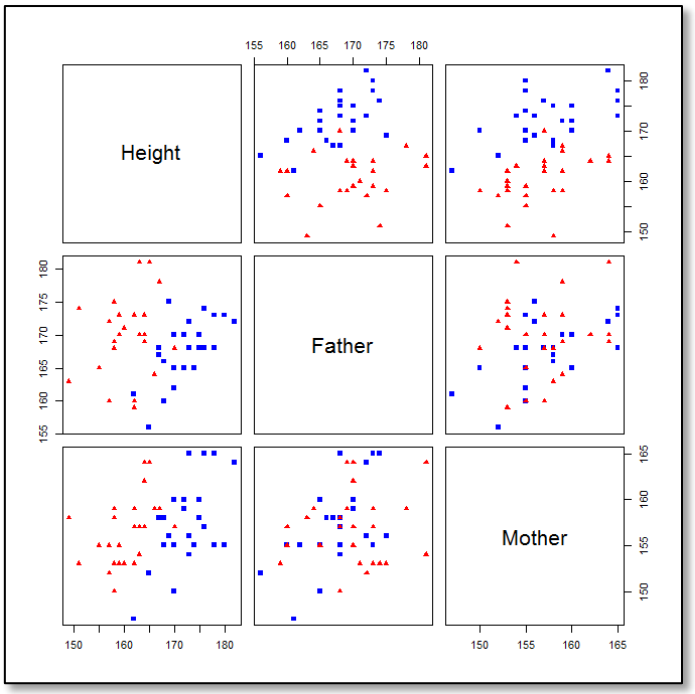


**図 1 男女別にマーカーを変更した散布図行列**

男女差のような質的情報を扱う場合，回帰係数の考え方，すなわちモデルの形を選ぶための戦略は，以下のようなものが考えられる。

1. 男女別に回帰モデルを計算する（切片や傾きが，男女で全く別と考える）。

[1] 説明変数が不足，線形モデルでは説明できない，など理由は種々考えられる。

2. 男女の身長差は切片の違いのみで，傾きは共通であるものとして計算する。
3. 男女の身長差は父母の影響（傾き）の違いであって，切片は共通であるものとして計算する。

**男女別に回帰モデルを計算**

この場合は，元のデータを男女別に分けて与える。 lm の呼び出しに subset 引数を追加する。まず，男性のみを取り出して分析するために，subset 引数に条件式 Sex==1 を与える。

```
> result2.1 = lm(Height ~ Father+Mother, data=data0, subset=Sex==1)
> summary(result2.1)

Call:
lm(formula = Height ~ Father + Mother, data = data0, subset = Sex ==
    1)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5221 -1.8259  0.3246  1.5626  6.2868

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.1664    30.2688   1.459   0.1587
Father        0.5047     0.2035   2.481   0.0212 *
Mother        0.2758     0.2087   1.322   0.1998
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.698 on 22 degrees of freedom
Multiple R-squared:  0.4517,    Adjusted R-squared:  0.4018
F-statistic: 9.061 on 2 and 22 DF,  p-value: 0.001347
```

同様に，女性のみを対象とするために，subset 引数に条件式 Sex==2 を与える。

```
> result2.2 = lm(Height ~ Father+Mother, data=data0, subset=Sex==2)
> summary(result2.2)

Call:
lm(formula = Height ~ Father + Mother, data = data0, subset = Sex ==
    2)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2789  -1.3013  -0.2504   2.3001   9.5442

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  51.7293    42.4202   1.219   0.2356
Father        0.1432     0.1544   0.928   0.3636
Mother        0.5393     0.2320   2.324   0.0298 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.4 on 22 degrees of freedom
Multiple R-squared:  0.2368,    Adjusted R-squared:  0.1674
F-statistic: 3.412 on 2 and 22 DF,  p-value: 0.05119
```

結果を要約すると，男性モデルでは，説明力は約 40%（決定係数が 0.4 ほど）であるが，Father のみが有意であるのに対して，女性モデルでは，説明力は 17%ほどしかなく，Mother のみが有意である，と言える。

一般的には，決定係数は過半数の 0.6（60%）以上でないと有効なモデルとは言えないので，男女別のモデルでも説明力が十分とは言えない。

**男女差を切片の違いと見た計算**

次のモデルでは，男女差を切片の違いと見る（しかし，傾きは共通である）。一般には，男女差をダミー変数と呼ぶ，人工的なデータに置き換えて計算できるが，R 言語では，該当する質的情報を factor と呼ぶデータタイプで表現しておくと効率が良い。このため，一般の数値情報である Sex を，factor に変換した Sex.f という列をデータフレームに追加する。

```
> data0$Sex.f = factor( data0$Sex, levels=1:2, labels=c("Male","Female") )
> data0[20:30,]
   No Height Sex Father Mother  Sex.f
20 20    176   1    174    165   Male
21 21    176   1    168    157   Male
22 22    178   1    173    165   Male
23 23    178   1    168    155   Male
24 24    180   1    173    155   Male
25 25    182   1    172    164   Male
26 26    149   2    163    158 Female
27 27    151   2    174    153 Female
28 28    155   2    165    155 Female
29 29    157   2    160    155 Female
30 30    157   2    172    152 Female
```

このように加工したデータフレームを用いると，基本モデルの数式を， Height ~ Father + Mother + Sex.f のように性別の情報を説明変数として追加するだけでよい。

```
> result3 = lm(Height ~ Father+Mother+Sex.f, data=data0)
> summary(result3)

Call:
lm(formula = Height ~ Father + Mother + Sex.f, data = data0)

Residuals:
    Min      1Q  Median      3Q     Max
-10.540  -2.014  -0.232   2.307   9.729

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  57.3254    25.0451   2.289  0.02673 *
Father        0.2408     0.1181   2.039  0.04726 *
Mother        0.4734     0.1480   3.198  0.00251 **
Sex.fFemale -11.8225     1.1853  -9.974 4.42e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.062 on 46 degrees of freedom
Multiple R-squared:  0.7255,    Adjusted R-squared:  0.7076
F-statistic: 40.53 on 3 and 46 DF,  p-value: 5.743e-13
```

この結果，Sex.f が Female に該当する時は，(全体の定数項 57.3254 から) 一律に 11.8225 低い値を取る（定数項は，男性の場合 57.3254 であるが，女性の場合 45.5029 となる)。また，モデル全体の説明力は 70%ほどとなり，各回帰係数も有意となっている。

**男女の身長差を傾きの違いと見た計算**

最後に，質的情報を傾きの違いとして表現するために，基本モデルの数式に， Height ~ (Father + Mother):Sex.f のように性別の情報を 「:」 で区切って追加する。

```
> result4 = lm(Height ~ (Father+Mother):Sex.f, data=data0)
> summary(result4)

Call:
lm(formula = Height ~ (Father + Mother):Sex.f, data = data0)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2433  -1.3899   0.0168   2.2593   9.5570

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         47.3343    25.0820   1.887  0.06560 .
Father:Sex.fMale     0.4944     0.2100   2.354  0.02298 *
Father:Sex.fFemale   0.1516     0.1294   1.172  0.24744
```

```
Mother:Sex.fMale      0.2667      0.2184    1.221   0.22834
Mother:Sex.fFemale    0.5582      0.1694    3.296   0.00192 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.02 on 45 degrees of freedom
Multiple R-squared:  0.7371,    Adjusted R-squared:  0.7137
F-statistic: 31.54 on 4 and 45 DF,  p-value: 1.551e-12
```

この結果，説明力は若干上昇して，約 71%となったが，女性の場合の Father からの影響，男性の場合の Mother からの影響が，有意ではなくなっている。

総合すると，男女差を切片の違いとして計算したモデルが，最も妥当なものと言える。