ロジスティック回帰分析の標準化偏回帰係数

青木繁伸

2012 年 6 月 8 日

ロジスティック回帰分析において，標準化偏回帰係数を出力しない統計解析ソフトで標準化偏回帰係数を求めるためにどうしたらよいかについて記述する。

以下のようなデータを使う。

```
> x1 <- c(1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0,
+ 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0,
+ 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0,
+ 1, 0, 0, 0, 0)
> x2 <- c(0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1,
+ 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0,
+ 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0,
+ 1, 0, 0, 0, 0)
> x3 <- c(0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1,
+ 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1,
+ 1, 0, 1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0, 0,
+ 0, 1, 0, 0, 0)
> x4 <- c(50.7, 47.4, 40.1, 61.5, 60.3, 51.2, 41.8, 59.7, 38.2, 51.8,
+ 57.1, 51.6, 47.6, 35.1, 61, 56.3, 49.4, 62.1, 44.5, 44.7, 51,
+ 56.4, 48.7, 44.9, 66.1, 49.9, 47, 60.3, 53.3, 43.2, 57.8, 57.3,
+ 41.5, 49.7, 33.5, 49.5, 57.3, 48.2, 44.6, 41.2, 60.9, 39.7, 64.8,
+ 64.7, 57.8, 49.6, 57.2, 41.2, 51.3, 44.2)
> y <- c(1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 0,
+ 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1,
+ 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1,
+ 0, 0, 1)
```

ダミー変数 $x1, x2, x3$ は factor とせずにこのまま数値データとして使ってよい（全く同じ結果になる）。

```
> d <- data.frame(x1, x2, x3, x4, y)
> a <- glm(y~x1+x2+x3+x4, data=d, family=binomial)
> summary(a)
Call:
glm(formula = y ~ x1 + x2 + x3 + x4, family = binomial, data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8289  -0.8437   0.2585   0.6145   1.8670
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.35728    2.65180  -2.020   0.0434 *
x1           0.94839    0.75398   1.258   0.2084
x2          -2.30690    0.96504  -2.390   0.0168 *
x3           2.56971    1.00720   2.551   0.0107 *
x4           0.10414    0.05164   2.017   0.0437 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 67.301  on 49  degrees of freedom
Residual deviance: 46.882  on 45  degrees of freedom
AIC: 56.882

Number of Fisher Scoring iterations: 5
> # d <- data.frame(x1=factor(x1), x2=factor(x2), x3=factor(x3), x4, y)
> # a <- glm(y~x1+x2+x3+x4, data=d, family=binomial)
> # summary(a)
```

ダミー変数 $x1, x2, x3$ と普通の変数 $x4$ を標準化してデータフレームを作る。平均値=0, 標準偏差=1 になる。

```
> d2 <- data.frame(x1=scale(x1), x2=scale(x2), x3=scale(x3), x4=scale(x4), y)
> sapply(d2[1:4], mean)
           x1            x2            x3            x4
 0.000000e+00 -6.664157e-17 -2.659765e-17 -3.747610e-16
> sapply(d2[1:4], sd)
x1 x2 x3 x4
 1  1  1  1
```

標準化したデータフレームを使ってロジスティック回帰分析を行う。表示される偏回帰係数が「標準化偏回帰係数」である。偏回帰係数以外の結果は全く同じであることを確認しておこう。

```
> b <- glm(y~x1+x2+x3+x4, data=d2, family=binomial)
> summary(b)
Call:
glm(formula = y ~ x1 + x2 + x3 + x4, family = binomial, data = d2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8289  -0.8437   0.2585   0.6145   1.8670

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.7018     0.3965   1.770   0.0767 .
```

```
x1            0.4790     0.3808   1.258   0.2084
x2           -1.0679     0.4467  -2.390   0.0168 *
x3            1.2600     0.4938   2.551   0.0107 *
x4            0.8499     0.4215   2.017   0.0437 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 67.301  on 49  degrees of freedom
Residual deviance: 46.882  on 45  degrees of freedom
AIC: 56.882

Number of Fisher Scoring iterations: 5
```

さて，重回帰分析の標準化偏回帰係数がどのようにして計算されるかを参考にすると，「標準化偏回帰係数」は，偏回帰係数に独立変数の標準偏差を掛けたものになっていることがわかる。実際に計算して，同じになることを確認しよう。

```
> coefficients(a)[-1] * sapply(d[1:4], sd)
        x1         x2         x3         x4
 0.4790082 -1.0678883  1.2599633  0.8499303
> all.equal(coefficients(a)[-1] * sapply(d[1:4], sd), coefficients(b)[-1])
[1] TRUE
```

よって，標準化偏回帰係数を出力しない統計解析ソフトで，標準化偏回帰係数を計算するには，表示された偏回帰係数に独立変数の標準偏差を掛ければよいことが示された。