R practice: Probit analysis (Source: R-intro-1.7.0 manual)

Minato Nakazawa

16 May 2011

1 Explantion in English

Consider a small, artificial example, from Silvey (1970).

On the Aegean island of Kalythos the male inhabitants suffer from a congenital eye disease, the effects of which become more marked with increasing age. Samples of islander males of various ages were tested for blindness and the results recorded. The data is shown below:

Age:	20	35	45	55	70
No. tested:	50	50	50	50	50
No. blind:	6	17	26	37	44

The problem we consider is to fit both logistic and probit models to this data, and to estimate for each model the LD50, that is the age at which the chance of blindness for a male inhabitant is 50%.

If y is the number of blind at age x and n the number tested, both models have the form

$$y \sim \mathrm{B}(n, F(\beta_0 + \beta_1 x))$$

where for the probit case,

$$F(z) = \Phi(z)$$

is the standard normal distribution function, and in the logit case (the default),

$$F(z) = e^z/(1 + e^z)$$

In both cases the LD50 is

$$\mathrm{LD50} = -\beta_0/\beta_1$$

that is, the point at which the argument of the distribution function is zero.

The first step is to set the data up as a data frame

```
> kalythos <- data.frame(x = c(20,35,45,55,70), n = rep(50,5),
                         y = c(6,17,26,37,44))
```

To fit a binomial model using `glm()` there are two possibilities for the response:

- If the response is a **vector** it is assumed to hold **binary** data, and so must be a 0/1 vector.
- If the response is a **two column matrix** it is assumed that the first column holds the number of successes for the trial and the second holds the number of failures.

Here we need the second of these conventions, so we add a matrix to our data frame:

```
> kalythos$Ymat <- cbind(kalythos$y, kalythos$n - kalythos$y)
```

To fit the models we use

```
> fmp <- glm(Ymat ~ x, family = binomial(link=probit), data = kalythos)
> fml <- glm(Ymat ~ x, family = binomial, data = kalythos)
```

Since the logit link is the default the parameter may be omitted on the second call. To see the results of each fit we could use

```
> summary(fmp)
> summary(fml)
```

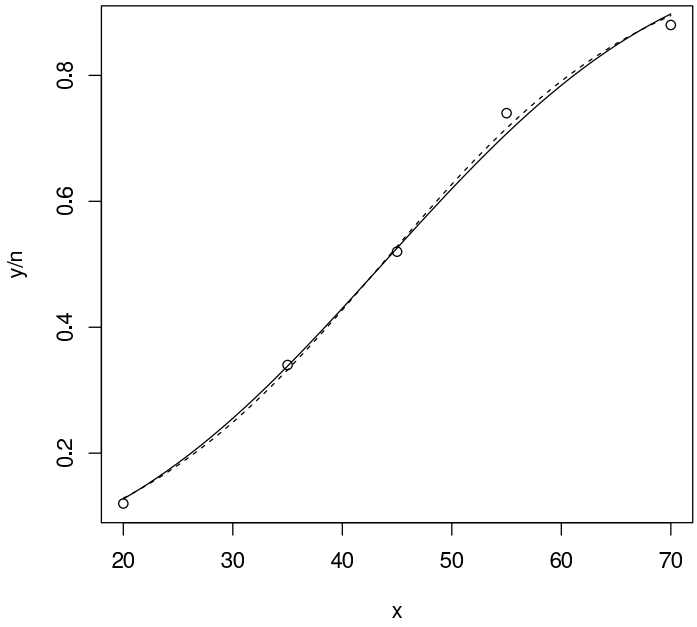
Both models fit (all too) well. To find the LD50 estimate we can use a simple function:

```
> ld50 <- function(b) -b[1]/b[2]
> ldp <- ld50(coef(fmp)); ldl <- ld50(coef(fml)); c(ldp, ldl)
```

The actual estimates from this data are 43.663 years and 43.601 years respectively.

To plot raw data and to draw fitted lines (probit in solid line, logit in dotted line), type as follows.

```
> plot(y/n ~ x, data=kalythos, type="p")
> ages <- data.frame(x=20:70)
> fitted.probit <- predict(fmp, ages, type="resp")
> fitted.logit <-  predict(fml, ages, type="resp")
> lines(ages$x,fitted.probit,lty=1)
> lines(ages$x,fitted.logit,lty=2)
```



2 日本語による解説（間瀬ら訳）

Silvey (1970) にある簡単な人工的な例を考えよう．

エーゲ海諸島のカリトス島の男性住民は，年齢とともに進行する，遺伝的な目の病気にかかる．さまざまな年齢の島の男性住民の盲目度が調査され，記録された．データは以下に示されている：

年齢:	20	35	45	55	70
被検者数:	50	50	50	50	50
盲目者数:	6	17	26	37	44

我々の考える問題は，このデータにロジスティックモデルとプロビットモデルを当てはめ，各モデルに対する LD50 値，つまり男性住民が盲目になる可能性が 50% を越える年齢を推定することである．

y を年齢 x での盲目者数，n を被検者数とすると，両方のモデルは

$$y \sim \mathrm{B}(n, F(\beta_0 + \beta_1 x))$$

の形になる．ここでプロビットモデルでは

$$F(z) = \Phi(z)$$

標準正規分布であり，ロジットモデル (既定) では

$$F(z) = e^z/(1 + e^z)$$

である．双方のモデルで LD50 値は

$$\mathrm{LD50} = -\beta_0/\beta_1$$

となる，つまり分布関数の引数がゼロとなるような点である．

最初のステップはデータをデータフレームとしてセットすることである

```
> kalythos <- data.frame(x = c(20,35,45,55,70), n = rep(50,5),
                         y = c(6,17,26,37,44))
```

glm() を用いて 2 項モデルを当てはめるには応答変数として 2 つの可能性がある：

- もし応答が「ベクトル」ならば，それは「バイナリ (2 値)」データを含んでいると仮定され，結局 0/1 ベクトルでなければならない．
- もし応答が「2 列の行列」ならば，最初の列は試行における成功の数，第 2 列は失敗の数を含んでいると仮定される．

ここでは，後者の場合を考えるので，データフレームに行列を加える：

```
> kalythos$Ymat <- cbind(kalythos$y, kalythos$n - kalythos$y)
```

モデルを当てはめるには次のようにする．

```
> fmp <- glm(Ymat ~ x, family = binomial(link=probit), data = kalythos)
> fml <- glm(Ymat ~ x, family = binomial, data = kalythos)
```

ロジット連結が既定値であるので，第 2 の呼出しではパラメータを省略しても良い．各当てはめの結果を見るには次のようにする．

```
> summary(fmp)
> summary(fml)
```

ともに (うますぎるほど) 良い当てはめを示す．LD50 値の推定値を見付けるには次のような簡単な関数を使う：

```
> ld50 <- function(b) -b[1]/b[2]
> ldp <- ld50(coef(fmp)); ldl <- ld50(coef(fml)); c(ldp, ldl)
```

このデータから得られる実際の推定値は，それぞれ 43.663 歳と 43.601 歳である．

生データ（年齢別盲目者割合）をプロットし，当てはめ曲線を描く（probit が実線，logit が破線）には，次のようにタイプする。

```
> plot(y/n ~ x, data=kalythos, type="p")
> ages <- data.frame(x=20:70)
> fitted.probit <- predict(fmp, ages, type="resp")
> fitted.logit <-  predict(fml, ages, type="resp")
> lines(ages$x,fitted.probit,lty=1)
> lines(ages$x,fitted.logit,lty=2)
```