#015★★☆☆ 2021.6.1 / 2023.7.24

ロジスティック解析とオッズ比。

臨床研究などで結果（アウトカム）が「副作用あり(1)」「副作用なし(0)」といった「カテゴリ変数（categorical variable)」で得られる場合があり、特に 2 種類の値のみをとるデータを「二値データ (binary data)」という。二値データに対しては 2 × 2 分割表を作成しカイ二乗検定やオッズ比を用いることが多いが、回帰分析として「ロジスティック回帰分析（logistic regression)」を用いる。カテゴリデータを従属変数として用いる場合には、そのまま用いるのではなくオッズの対数値（ロジット、logit）を用い、カテゴリデータに対してロジットを得ることをロジット変換（logit transform）という。

観測値	副作用の有無		
	有	無	合計
Drug X	250	75	325
Drug Y	155	120	275
合計	405	195	600

表のような分割表では要因はひとつ（Drug X か Drug Y か）であるが、複数の要因（独立変数）を考える場合の回帰モデル式は：

$$logit = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \beta_3 \times x_3 + \cdots$$

となる。ここで β_0 は回帰分析の切片、x_1、x_2 …は各独立変数、β_1、β_2 … はその回帰係数である。最も簡単な、ひとつだけ独立変数を含む場合は次のようになる。

$$logit = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times x_1$$

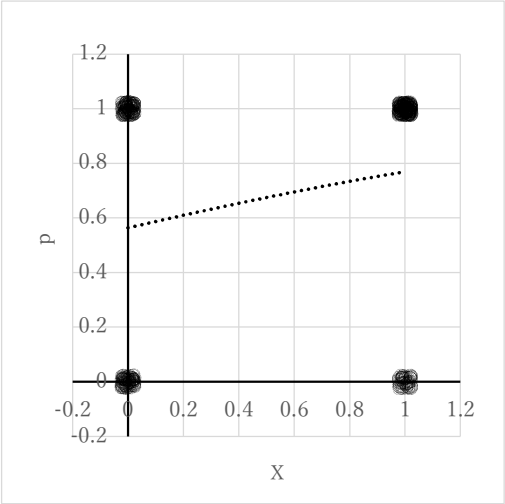
実際に上表にある数値を用いて計算を確認してみる。Drug X 投与時で x_1=1 とすると、そのときの副作用発症率は 250 / 325 = 0.769、Drug Y 投与時で x_1=0 とすると、そのときの副作用発症率は 155 / 275 = 0.564 なので、ln {0.769 / (1 - 0.769) } = 1.204 = $\beta_0 + \beta_1$、ln {0.564 / (1 - 0.564) } = 0.256 = β_0 となり、β_0 = 0.256、β_1 = 1.204 - 0.256 = 0.948 と得られる。

実際にはより複雑なモデルに対応するために統計解析フトウェアを利用して「最尤推定法」と呼ばれる手法で係数 β を求める。上記のデータについてプログラム言語 R を用いて計算した結果（右図)、β_0 = 0.256、β_1 = 0.948 と同じ値が得られている。

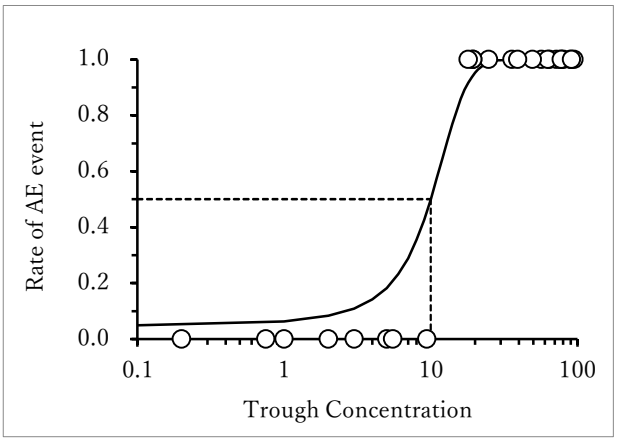
```
> results <- glm(data=df, formula = Y ~ X, family = binomial)
> summary(results)

Call:
glm(formula = Y ~ X, family = binomial, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7125  -1.2878   0.7244   1.0708   1.0708

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.2559     0.1216   2.105   0.0353 *
X             0.9480     0.1792   5.290 1.22e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 756.7  on 599  degrees of freedom
Residual deviance: 727.9  on 598  degrees of freedom
AIC: 731.9

Number of Fisher Scoring iterations: 4
```

係数 β はその要因の OR との間に OR=exp(β) の関係があり、β_1 の場合 OR = exp(0.948) = 2.58 となる。また OR の 95%信頼区間は exp (OR±1.96×SE)（SE：standard (std) error）により得られ、下限 1.82、上限 3.67 と計算できる。ここで 1.96 は 95%区間を算出するときによく用いる数値である。

結果を図に示した。データ数が多いため個々のプロットが黒い塊となっている。点線は回帰分析の結果を x が 0 から 1 まで連続的に変化させたときの y (副作用発症) の確率 p の推移を表す。$x = 0$ のとき p = 0.564、$x = 1$ のとき p = 0.769 と分割表の要因 X、要因 Y それぞれの副作用発症率と一致する。



ロジスティック回帰分析は、独立変数が連続変数の場合でも適用できる。右下図は、横軸にある薬物の定常状態での血中濃度トラフ値（最低値）を、縦軸に副作用発症の有無を二値で示した架空データである。ロジスティック回帰分析を適用するとトラフ濃度と副作用発症有無との関係は図のようになる。縦軸は副作用発症確率ともみなすことができるので、例えば点線のように確率が 50%となるようなトラフ値は約 10 と読み取ることができ、目標トラフ濃度を設定することで投与計画に活かすことができる。



（参考）

観測値	結果事象発現の有無		
	有	無	合計
要因 X	A	B	A+B
要因 B	C	D	C+D
合計	A+C	B+D	A+B+C+D

上表に示す分割表の記号 A～D を用いると、$\ln(A/B) = \beta_0 + \beta_1 \times 1$、$\ln(C/D) = \beta_0 + \beta_1 \times 0$ より、$\beta_1 = \ln(A/B) - \ln(C/D) = \ln\{(A/B)/(C/D)\} = \ln(AD/BC) = \ln(OR)$ となり、OR=exp(β_1) の関係が説明できた。