

ロジスティック回帰分析（目的変数の確率値は0か1に限定）

1. 目的

サンプルデータの目的変数の確率値が 0 もしくは 1 に限定して指定されている場合に利用します。

モデル式は以下で、説明変数は $x_1 \sim x_m$ です。

$$P = \frac{1}{1 + \exp(- (a_0 + a_1x_1 + a_2x_2 + \dots + a_mx_m))}$$

入力データのサンプルから 最良のパラメータ $a_0 \sim a_m$ を推測し、新たにデータが与えられた場合の 確率 P を 推定します。

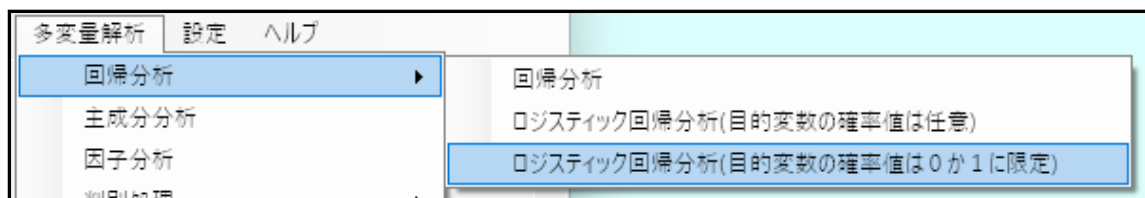
「考え方」で説明しているように、 $P=0$ または $P=1$ の場合、上式をそのまま利用しての計算アルゴリズム（最小自乗法）は適応できない為、ここでは別のアルゴリズム（最尤法）により計算します。

なお、P が (0.0, 1.0) の範囲で任意の値の場合には、この機能ではなく、ロジスティック回帰分析（目的変数の確率値は任意） を利用してください。

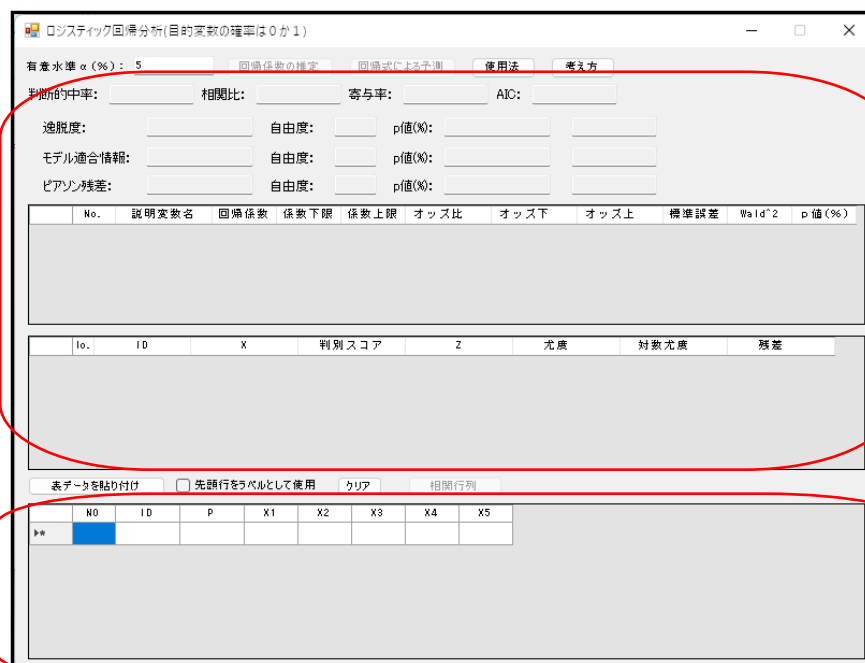
2. 使用法

(1) メニューの選択

メニュー「多変量解析→ロジスティック回帰分析(目的変数の確率値は0か1に限定)」を選択します。



(2) パネルが表示されます。

A screenshot of the 'ロジスティック回帰分析(目的変数の確率値は0か1)' (Logistic Regression Analysis (Probability of the dependent variable is 0 or 1)) panel. The panel has several tabs: '有意水準 α (%) : 5', '回帰係数の推定', '回帰式による予測', '使用法', and '考え方'. The '使用法' (Usage) tab is selected. Below the tabs are input fields for '判断的中率:', '相関比:', '寄与率:', and 'AIC:'. There are also input fields for '逸脱度:', '自由度:', 'p値(%)', 'モデル適合情報:', '自由度:', 'p値(%)', and 'ピアソン残差:', '自由度:', 'p値(%)'. Below these are two tables. The first table has columns: 'No.', '説明変数名', '回帰係数', '係数下限', '係数上限', 'オッズ比', 'オッズ下', 'オッズ上', '標準誤差', 'Wald*2', and 'p値(%)'. The second table has columns: 'ID', 'X', '判別スコア', 'Z', '尤度', '対数尤度', and '残差'. At the bottom, there are checkboxes for '表データを貼り付け' (Paste table data), '先頭行をラベルとして使用' (Use the first row as a label), and 'クリア' (Clear). There is also a '相関行列' (Correlation matrix) button. A red line highlights the entire panel area. A red line also highlights the bottom section of the panel, which contains the data input table.

計算結果が
表示される部分

データを
入力する部分

(3) 計算対象データを入力

不整脈有無についてのデータを例題に説明します。
表計算ソフトに計算対象データを以下のように定義します。

ID	不整脈有無	喫煙有無	飲酒有無	ギャンブル嗜好
ID-1	1	1	1	1
ID-2	1	1	1	1
ID-3	1	1	1	1
ID-4	1	1	0	1
ID-5	1	1	0	1
ID-6	1	1	0	0
ID-7	1	0	1	0
ID-8	1	0	0	0
ID-9	0	1	0	1
ID-10	0	1	0	0
ID-11	0	0	1	0
ID-12	0	0	1	0
ID-13	0	0	0	1
ID-14	0	0	0	0
ID-15	0	0	0	0
ID-16	0	0	0	0
ID-17	0	0	0	0
ID-18	0	0	0	0
ID-19	0	0	0	0
ID-20	0	0	0	0

最初の列は ID を指定します。

2 番目の列は 目的変数で、ここには 0 または 1 が指定されます。
ここでの目的変数は不整脈の有無で、0 は確率値 0、1 は確率値 1 を示します。

3 番目以降の列は説明変数で、2 値のカテゴリ値、または数量が指定できます。
ここでの説明変数は、

- ・喫煙の有無： 0：なし、1：有り
- ・飲酒の有無： 0：なし、1：有り
- ・ギャンブル嗜好： 0：なし、1：有り

であり、ここでは いずれも 2 値のカテゴリ値となっています。
カテゴリ値の場合は、2 値（0 または 1）に限ります。

ここでは不整脈になる確率を扱っているので、

- ・喫煙していると確率が高そうなので 1 を指定
- ・喫煙してないと確率が低そうなので 0 を指定

としているわけですが、逆に指定しても計算はできます。

ただ、結果の解釈が やや わかりづらくなると思います。

また 設定値も（0，1）でなく任意の 2 値でも計算はできますが、
やはりわかりづらくなるので、（0，1）を指定するのが無難でしょう。

このことは 飲酒やギャンブルについても 同様です。

また、説明変数には、2 値のカテゴリ値 と 普通の数量 とを混せて使用することも可能です。

表計算上の 以下の赤い部分を 計算に利用します。

ID	不整脈有無	喫煙有無	飲酒有無	ギャンブル嗜好
ID-1	1	1	1	1
ID-2	1	1	1	1
ID-3	1	1	1	1
...
ID-18	0	0	0	0
ID-19	0	0	0	0
ID-20	0	0	0	0

赤い部分を コピーして、 表データを貼り付け を
クリックすると、入力が完了します。

表データを貼り付け <input checked="" type="checkbox"/> 先頭行をラベルとして使用 クリア 相						
	NO	ID	不整脈有無	喫煙有無	飲酒有無	ギャンブル嗜好
▶	1	ID-1	1	1	1	1
	2	ID-2	1	1	1	1
	3	ID-3	1	1	1	1
	4	ID-4	1	1	0	1
	5	ID-5	1	1	0	1
	6	ID-6	1	1	0	0

(4) 計算を実行

「回帰係数の推定」 ボタンを押すことで、計算結果が表示されます。

判断的中率: 0.8 相関比: 0.422595 寄与率: 0.355828 AIC: 25.34141											
逸脱度:		17.34141	自由度:	16	p値(%):	36.3844	有意でない				
モデル適合情報:		9.579056	自由度:	3	p値(%):	2.250494	有意				
ピアソン残差:		19.7449	自由度:	16	p値(%):	23.19308	有意				
	No.	説明変数名	回帰係数	係数下限	係数上限	オッズ比	オッズ下	オッズ上	標準誤差	Wald^2	p 値 (%)
▶	0	定数項	-2.374316	0	0	0	0	0	1.081906	4.816123	2.81948
	1	喫煙有無	2.642819	-0.417829	5.703467	14.05276	0.6584749	299.9054	1.561198	2.865622	9.04907
	2	飲酒有無	1.94843	-0.78448...	4.68135	7.017664	0.4563526	107.9157	1.394028	1.95356	16.2203

	No.	ID	X	判別スコア	Z	尤度	対数尤度	残差
▶	1	ID-1	2.720459	0.9382231	-0.05796049	0.9382231	-0.06376748	0.06584454
	2	ID-2	2.720459	0.9382231	-0.05796049	0.9382231	-0.06376748	0.06584454
	3	ID-3	2.720459	0.9382231	-0.05796049	0.9382231	-0.06376748	0.06584454
	4	ID-4	0.7720284	0.6839595	-0.2161589	0.6839595	-0.3798566	0.4620748
	5	ID-5	0.7720284	0.6839595	-0.2161589	0.6839595	-0.3798566	0.4620748

(5) 計算結果の説明

回帰係数等について、以下の結果が得られています。

説明変数名	回帰係数	オッズ比	標準誤差	Wald^2	p値 (%)
定数項	-2.3743	0.0000	1.0819	4.8161	2.8195
喫煙有無	2.6428	14.0528	1.5612	2.8656	9.0491
飲酒有無	1.9484	7.0177	1.3940	1.9536	16.2204
ギャンブル嗜好	0.5035	1.6545	1.5545	0.1049	74.6009

まず、回帰係数は 以下ようになります。

説明変数名	回帰係数
定数項	-2.3743
喫煙有無	2.6428
飲酒有無	1.9484
ギャンブル嗜好	0.5035

これは 以下の式

$$P = \frac{1}{1 + \exp \left(- \left(a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_m x_m \right) \right)}$$

において

$a_0 = -2.3743$
 $a_1 = 2.6428$
 $a_2 = 1.9484$
 $a_3 = 0.5035$

(定数項)

(喫煙有無)

(飲酒有無)

(ギャンブル嗜好)

であることを示します。
これを用いて、新たな $x_1 \sim x_m$ が指定された時、上式より確率Pが求まります。

続いて オッズ比 です。 計算結果は以下です。

説明変数名	オッズ比
定数項	0.0000
喫煙有無	14.0528
飲酒有無	7.0177
ギャンブル嗜好	1.6545

ここで オッズ、 オッズ比 について 少し説明を加えます。
ある事象が起こる確率を p とすると、起こらない確率は $1 - p$ で、
オッズは $p / (1 - p)$ と定義されます。

オッズ比は2つのオッズの比で定義されますが、入力データを例にオッズ、
オッズ比について説明します。

今検討している、不整脈について、要因と考えられる

- ・喫煙
- ・飲酒
- ・ギャンブル嗜好

に關しての クロス表を入力データから作成すると以下となります。
表のセルの数字は、該当条件に適う人数です。

	不整脈	
	あり	なし
喫煙してる	6	2
喫煙してない	2	10

	不整脈	
	あり	なし
飲酒してる	4	2
飲酒してない	4	10

	不整脈	
	あり	なし
ギャンブル好き	5	2
ギャンブル嫌い	3	10

不整脈と喫煙の關係についての オッズ および オッズ比 を考えます。

喫煙してる場合、不整脈ありの率を p_1 、なしの確率を $1 - p_1$ 。

喫煙してない場合、不整脈ありの確率を p_2 、なしの確率を $1 - p_2$ 。

とすると、

- ・喫煙してる場合のオッズ $\frac{p_1}{1-p_1}$ は A / B であり 3
- ・喫煙してない場合のオッズ $\frac{p_2}{1-p_2}$ は C / D であり 0.2

となります。

	不整脈		
	あり	なし	オッズ
喫煙してる	A(=6)	B(=2)	A/B (=3)
喫煙してない	C(=2)	D(=10)	C/D (=0.2)
		オッズ比	AD/BC (=15)

不整脈と喫煙の関係が大きいと予想できるなら、上式の

- ・ 1 番目の喫煙してる場合のオッズ $\frac{p_1}{1-p_1} = A/B$ は 1 より大きいはず
- ・ 2 番目の喫煙してない場合のオッズ $\frac{p_2}{1-p_2} = C/D$ は 1 より小さいはず

と予想されます。

そして、2つのオッズの比がオッズ比であり、以下で定義されます。

$$\text{オッズ比} = (A/B) / (C/D)$$

このオッズ比が大きいほど、目的変数（不整脈）に対する説明変数（喫煙）の影響は大きく、逆に小さいほど影響は小さいと考えます。

例題のデータでの、不整脈と喫煙との オッズ比 は 1.5 です。

これは、喫煙者が不整脈になる割合は、非喫煙者が不整脈になる割合の 1.5 倍であるということで、喫煙すると不整脈のリスクは大きいと考える理由といえます。

なお、オッズ、オッズ比については、以下のように定義することもできます。

	不整脈		
	あり	なし	
喫煙してる	A (=6)	B (=2)	オッズ比
喫煙してない	C (=2)	D (=10)	
オッズ	A/C (=3)	B/D (=0.2)	
		AD/BC (=1.5)	

不整脈ありの場合、喫煙してる確率を p_1 、してない確率を $1 - p_1$ 、

不整脈なしの場合、喫煙してる確率を p_2 、してない確率を $1 - p_2$ 、

とすると、

- ・ 不整脈ありの場合のオッズは A/C で 3
- ・ 不整脈なしの場合のオッズは B/D で 0.2

となります。

不整脈と喫煙の関係が大きそうだと予想できるなら、上式の

- ・ 1 番目の不整脈ありの場合のオッズ $\frac{p_1}{1-p_1} = A/C$ は 1 より大きいはず
- ・ 2 番目の不整脈なしの場合のオッズ $\frac{p_2}{1-p_2} = B/D$ は 1 より小さいはず

と予想されます。

そして、2つのオッズの比がオッズ比であり、オッズ比 $= (A/C) / (B/D)$ と定義されますが、このオッズ比が大きいほど、目的変数に対する説明変数の影響は大きく、逆に小さいほど影響は小さい と考えます。

結局オッズ比は どちらも AD/BC で 結果は変わりません。

最初に説明したオッズの考え方は、不整脈となる確率 に注目して計算しており、

後で説明したオッズの考え方は、喫煙者となる確率 に注目して計算しています。

そして、最後のオッズ比で、双方の関係の強弱を同じ数量で評価している

ということになります。

次に、クロス表によれば、

- ・不整脈と喫煙のオッズ比は 1.5
- ・不整脈と飲酒のオッズ比は 5
- ・不整脈とギャンブルのオッズ比は 8.3

となります。

ですので、クロス表から

- ・不整脈への影響は、喫煙が一番大きい
- ・次にギャンブルの影響が大きい
- ・飲酒の影響が一番小さい

ということになります。

一方ロジスティック回帰分析では、

- ・不整脈と喫煙のオッズ比は 14.05
- ・不整脈と飲酒のオッズ比は 7.02
- ・不整脈とギャンブルのオッズ比は 1.65

となり、各々のオッズ比は異なり、また影響の順位も異なってます。

これについては、変量間の関係を考慮しなければなりません。

不整脈、喫煙、飲酒、ギャンブル の変量間での相関係数は以下です。

変数名	不整脈有無	喫煙有無	飲酒有無	ギャンブル嗜好
不整脈有無	1.0000	0.5833	0.3563	0.4708
喫煙有無	0.5833	1.0000	0.1336	0.6847
飲酒有無	0.3563	0.1336	1.0000	0.2059
ギャンブル嗜好	0.4708	0.6847	0.2059	1.0000

喫煙と飲酒の相関係数は 0.13 ですが、喫煙とギャンブルの相関係数は 0.68 であり お互いの影響が大きいことがわかります。

つまり、クロス表によるオッズ比には、相関関係による影響があるため、その影響を外したオッズ比を求める必要があります。

ロジスティック回帰分析でのオッズ比は、これを考慮したもので、不整脈とギャンブルの相関のうち、喫煙の影響を除去した“より正確な”オッズ比を示していると考えます。

ロジスティック回帰分析でのオッズ比については、「考え方」の方で説明していますので、そちらをご覧ください。

したがって、クロス表でのオッズ比は参考になりますが、説明変数の影響を考慮する際には、ロジスティック回帰分析でのオッズ比を重視します。

次に標準誤差です。

標準誤差は 偏回帰係数 $a_0 \sim a_3$ の標準誤差であり、普通の回帰分析の標準誤差と考え方は同じです。（回帰分析の項の標準誤差を参照）

“Wald²” はワルドの χ^2 値といい、偏回帰係数が 0 かどうかの検定統計量で偏回帰係数を標準誤差で割り、自乗したものです。

これは自由度 1 の χ^2 分布に従い、統計量から有意確率 p 値が計算されます。
この p 値が有意水準を下回れば、有意（偏回帰係数は 0 とはみなせない）、
逆に上回れば、有意でない（偏回帰係数は 0 とみなせる）ということです。

次に 各入力データについて 以下のように出力されています。

No.	ID	X	判別スコア	Z	尤度	対数尤度	残差
1	ID-1	2.7205	0.9382	-0.0580	0.9382	-0.0638	0.0658
2	ID-2	2.7205	0.9382	-0.0580	0.9382	-0.0638	0.0658
3	ID-3	2.7205	0.9382	-0.0580	0.9382	-0.0638	0.0658
4	ID-4	0.7720	0.6840	-0.2162	0.6840	-0.3799	0.4621
...
20	ID-20	-2.3743	0.0852	-0.0779	0.9148	-0.0890	0.0931

- ・ X は $X = a_0 + a_1x_1 + a_2x_2 + a_3x_3$ 。
- ・ 判別スコアは確率値のことで、 $Y = \frac{1}{1+e^{-x}}$ 。
- ・ Z 値は $Z = -Y * (1 - Y)$ 。
- ・ 尤度は目的変数の入力値が 1 の時は Y 、入力値が 0 の時は $1 - Y$ 。
- ・ 対数尤度は 上記の尤度の自然対数。
- ・ 残差は $(P - Y) * (P - Y) / Y / (1 - Y)$ （P は入力値の 1 または 0）。

となります。

判断的中率:	0.8	相関比:	0.422595	寄与率:	0.355828	AIC:	25.34141
逸脱度:	17.34141	自由度:	16	p値(%):	36.3844		有意でない
モデル適合情報:	9.579056	自由度:	3	p値(%):	2.250494		有意
ピアソン残差:	19.7449	自由度:	16	p値(%):	23.19308		有意

判断的中率は、

- ・ 判別スコアが 0.5以上なら 1
- ・ 判別スコアが 0.5未満なら 0

として、入力値の 1, 0 と比較した場合の 正解率 です。

例題の場合 0.8 となっています。

相関比は、判別スコアから計算される群間変動を全体変動で割ったもので 1に近ければ、予測精度は高いとみなします。

AIC は モデルの適合度を表す指標で、各入力データについて計算された対数尤度の合計 を LL として、

$$\cdot AIC = -2 * LL + 2 * (\text{説明変数の数} + 1)$$

と定義されます。右式の第 1 項は、モデルの当てはまり度合い、第 2 項は変数が増えることによるペナルティで、小さい方が望ましいとされます。

例題の場合

$$AIC = -2 * (-8.67) + 2 * (3 + 1) = 25.34$$

となります。

逸脱度は 上記の $-2 * LL$ で定義され、この値は大きいほど、当てはまりが悪いと考えられます。

これは 自由度 = データ数 - 説明変数の数 - 1 の χ^2 分布に従い、統計量から有意確率 p 値が計算されます。

この p 値が有意水準を下回れば有意、上回れば有意でないということです。

入力データで、目的変数の値が1である数を n_1 、0である数を n_2 、

入力データ数 $n (=n_1 + n_2)$ として、

$$K_{dat} = n_1 * \log(n_1) + n_2 * \log(n_2) - n * \log(n) \text{ とします。}$$

寄与率は $(-2 * K_{dat} - \text{逸脱度}) / (-2 * K_{dat})$ と定義されます。

寄与率が高いほど、判別精度は高いことになります。

また、モデル適合情報は、 $-\text{逸脱度} - 2 * K_{dat}$ と定義され、

自由度が説明変数の数（例題では 3）の χ^2 分布に従います。

ピアソン残差は、各入力データについて計算された残差の合計です。

これは 自由度 = データ数 - 説明変数の数 - 1 の χ^2 分布に従い、統計量から有意確率 p 値が計算されます。

この p 値が有意水準を 上回れば 有意、下回れば 有意ではありません。

（ピアソン残差の検定は、通常と逆で 検定量 > p 値 の時有意です）

(6) 回帰式による予測

回帰式による予測

を押すことで、推計を実行します。

NO	説明変数名	値
1	喫煙有無	-
2	飲酒有無	-
3	ギャンブル嗜好	-

計算 確率:

説明変数の

- ・喫煙有無
- ・飲酒有無
- ・ギャンブル嗜好

に 値(この場合は 0 または 1)を代入して、確率を 計算できます。

NO	説明変数名	値
1	喫煙有無	1
2	飲酒有無	1
3	ギャンブル嗜好	0

計算 確率: 0.9017539

(7) 変数を計算対象から外す

入力データによっては、説明変数が回帰分析において有意でない場合があります。

その場合、説明変数を省いて 再計算することができます。

通常の回帰分析と同様です。