Logistic 回帰モデルにおける Wald 検定と TypeⅢ型検定

株式会社バイオスタティスティカル　リサーチ
古川敏仁

Logistic 回帰モデルのパラメータ推定値βへは最尤法により推定される。最尤法では、対数尤度関数の漸近正規性より、パラメータ推定値の分散（標準誤差）が推定される。

今、p 個の説明変数　X=(x1 x2 ・・・xp)を考え、ロジット g(X)が下記のパラメータの線形式で示されるとする（Applied Logistic Regression second edition, Daivid W. Hosmer　p31）。

$$g(X)=\beta 0+\beta 1\,x1+\beta 2\,x2+\ldots+\beta p\,xp \quad (2.1)$$

このとき、βj の推定値βへj の統計学的有意性は、βj の正規性を利用して、正規検定で簡単に行うことができる。

すなわち、Wj が 1.96 を超えれば、片側 5%で有意となる。これを Wald 検定と言う。

$$Wj=\frac{\widehat{\beta j}}{S\widehat{E}(\widehat{\beta j})}$$

例えば、LOEBWT の RACE 変数をカテゴリとした、SAS 出力は以下のようになる。

```
PROC LOGISTIC DATA= LOEBWT;
 CLASS RACE;
 MODEL LOW= RACE /COBV;
RUN;
```

最大尤度推定値の分析						
パラメータ		自由度	推定値	標準誤差	Wald カイ 2 乗	Pr > ChiSq
Intercept		1	0.6612	0.1759	14.1255	0.0002
RACE	1	1	0.4936	0.2236	4.8716	0.0273
RACE	2	1	−0.3511	0.2889	1.4771	0.2242

RACE(1)に関しては、w=0.4936÷0.223=2.21

SASでは、正規検定の代わりに、wjを自乗した自由度 1 のχ自乗検定の形で示されているためw^2=2.21^2=4.87 となっている。

$$Wj^2 = \left(\frac{\hat{\beta}j}{S\hat{E}(\hat{\beta}j)} \right)^2$$

なぜ、正規検定の代わりに、χ自乗検定の形で示しといるかというと、それは、多変量への拡張性が良いからである。

今、個々のパラメータ推定値βjの推定値β^jの検定について示した。しかし、RACEは白人、黒人、その他の3つのカテゴリからなる変数であり、偶然RACE(1)＝黒人だけが、集団平均から離れている可能性がある（SASのLogisticプロシジャでは、CLASSで指定された変数は　deviation from means coding :Effect codingであり、その解釈には注意が必要である）。RACE変数全体として、有意な情報を持っているかの検定も重要となる。これがTypeⅢ型の検定である。先ほどのプログラムの出力は、自由度2でp=0.0854で、RACEは有意とはなっていない。

Wald カイ 2 乗	自由度	Pr > ChiSq
4.9209	2	0.0854

このχ自乗値はどのように計算したのであろうか。もし、p個のパラメータが独立であれば、そのWj²の総和は自由度pのχ自乗分布に従う。

ためしに、RACE(1)とRACE(2)のχ自乗値を足してみると、4.8716+1.4771≒5.35≠4.92となり、TypeⅢ型のχ自乗値とは一致しない。これは、RACE(1)、RACE(2)が背反なカテゴリ変数であり集団平均からの差を求めるため、β^1=0.4936とβ^2=−0.3511とは独立ではないことを示している。

Applied Logistic Regression second edition, Daivid W. Hosmer　p39では、p個の変数の独立成分のWaldχ自乗統計量の総和（p自由度）を求める式が書かれている。

$$W = \hat{\beta}' \left[\widehat{Var}(\hat{\beta}) \right]^{-1} \hat{\beta} = \hat{\beta}' (X'VX) \hat{\beta}$$

このような、行列表現は、最もシンプルな場合、つまりp=2で考えると理解しやすい。
今、Var(β^)の逆行列をφ11、φ12、φ21、φ22とおいてやって、実際の計算は何をやっているのかと確認すると

$$W = \hat{\beta}' \left[\widehat{Var}(\hat{\beta}) \right]^{-1} \hat{\beta} = \begin{bmatrix} \beta 1 & \beta 2 \end{bmatrix} \begin{bmatrix} \phi 11 & \phi 12 \\ \phi 21 & \phi 22 \end{bmatrix} \begin{bmatrix} \beta 1 \\ \beta 2 \end{bmatrix} = \begin{bmatrix} \beta 1\phi 11 + \beta 2\phi 21, \beta 1\phi 12 + \beta 2\phi 22 \end{bmatrix} \begin{bmatrix} \beta 1 \\ \beta 2 \end{bmatrix}$$

$$= \beta 1\beta 1\phi 11 + \beta 1\beta 2\phi 21 + \beta 1\beta 2\phi 12 + \beta 2\beta 2\phi 22$$

もし共分散がなければ

$$= \frac{\beta 1^2}{Var(\beta 1)} + \frac{\beta 2^2}{Var(\beta 2)}$$

すなわち、もし、β1 とβ2 が独立（共分散がなければ）であれば、個々のパラメータ推定値の Wald χ 自乗の和を求めることになる。ただ、現実はそうではないので、共分散分だけ調整した W 統計量を求めることになる。共分散が存在すれば以下を計算することになる。

$$W = \frac{Var(\beta 2)\beta 1^2}{(Var(\beta 1)Var(\beta 2) - Cov^2(\beta 1, \beta 2))} + \frac{Var(\beta 1)\beta 2^2}{(Var(\beta 1)Var(\beta 2) - Cov^2(\beta 1, \beta 2))} + \frac{-Cov(\beta 1, \beta 2))\beta 1\beta 1}{(Var(\beta 1)Var(\beta 2) - Cov^2(\beta 1, \beta 2))} \quad (1)$$

この手のことは、実際に計算してみると理解が早い。SAS によるパラメータ推定値の分散共分散行列は下記である。

推定共分散行列			
パラメータ	Intercept	RACE1	RACE2
Intercept	0.03095	−0.01189	0.021574
RACE1	−0.01189	0.050008	−0.04063
RACE2	0.021574	−0.04063	0.083475

RACE の TypeⅢ型検定に必要な分散共分散行列は、その中の下記である。

パラメータ	RACE1	RACE2
RACE1	0.050008	−0.04063
RACE2	−0.04063	0.083475

また、必要なパラメータ推定値は下記である。

パラメータ		β 推定値
RACE	1	0.4936
RACE	2	−0.3511

これを、EXCEL で(1)に入れると W=4.92 となる。

また、SAS には IML という行列計算言語があり、この手の計算がわかりやすく計算できる。下記のように、それぞれ行列に実際の値を入れ、V の逆行列を INV 関数で求めると、以下のような出力が得られる。

```
PROC  ML;
RESET PRINT;
```

```
RESET LOG;

b={0.4936 -0.3511};
 v={0.050008    -0.04063,-0.04063 0.083475};
 gv=inv(v);
 w=b*gv*b`; ;
```

SAS システム

B	1 row	2 cols	(numeric)
	0.4936		-0.3511

V	2 rows	2 cols	(numeric)
	0.050008		-0.04063
	-0.04063		0.083475

GV	2 rows	2 cols	(numeric)
	33.077472		16.099883
	16.099883		19.815972

W	1 row	1 col	(numeric)
			4.9214513

まとめ

- Logistic 回帰モデルの TypeⅢ型検定は Wald 統計量を用いている。
- Wald 統計量はパラメータ推定値が正規分布することを利用している。
- Wald χ 自乗値は、個々のパラメータに関する χ 自乗値を独立になるように調整し、その総和を求めたものである。ゆえに、p 個の Wald χ 自乗値の自由度は p である。
- $W = \hat{\beta}'\left[\widehat{Var}(\hat{\beta})\right]^{-1}\hat{\beta} = \hat{\beta}'(X'VX)\hat{\beta}$ のような表現は一見難しそうに見えるが、行列は 2×2、共分散なしで考えると、計算の意味が理解しやすい。
- また、今回の例のように、実際に計算すれば、案外、慣れるのは早い。この行列式が理解できるだけでも、SAS などのマニュアルのかなりの部分が理解可能となる。分からなくなったら何度も計算することが理解のためには重要である。