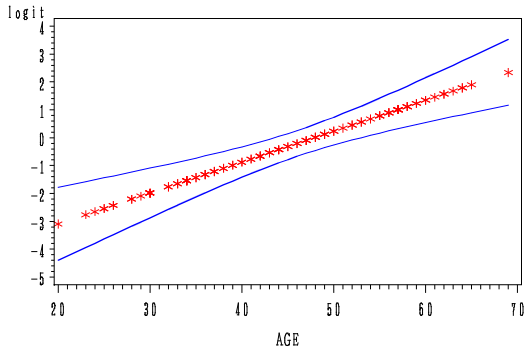
連続量の説明変数xによる単変量logistic回帰のロジット推定量$\hat{g}(x)$の信頼区間を最小にするxの値

西野、金、古川

目的:　図1はCHDデータにおけるCHD発症の有無と説明変数年齢(x)との関係をlogistic回帰で推定した場合のロジット推定量$\hat{g}(x)$の95%信頼区間である。この信頼区間は年齢平均 44.38 歳近辺で最も信頼区間幅が狭くなっているように見える。しかし、教科書にあるように実は、最も狭いところは説明変数の平均値近辺ではあるが平均値には一致しない。では単変量 Logistic model でロジット推定量$\hat{g}(x)$の信頼区間を最小にする変数の値 *x0* は、どのような値であるか確認してみよう。

図1　年齢とLogit推定値、95%信頼区間



ロジット推定量のx=xiにおける分散は、式(1.18)（Chapter1 page19）で求められ、xの関数となる。

ロジット推定量の分散はxの関数なので、それを最小にするxの値は、式(1.18)を微分して0とおいた(1)から、(2)式となる(第 2 回 ALR 勉強会 金さん班資料を参照)。つまり、CHDデータの場合は、年齢平均 44.38 歳ではなく、46.08 歳が最も$\hat{g}(x)$の推定精度の良い点となる。

$$\hat{\mathrm{var}}[\hat{g}(x)] = \hat{\mathrm{var}}(\hat{\beta}_0) + x^2\,\hat{\mathrm{var}}(\hat{\beta}_1) + 2x\,\hat{\mathrm{cov}}(\hat{\beta}_0, \hat{\beta}_1) \qquad (1.18)$$

$$\frac{\partial}{\partial x} Var(g(x)) = \frac{\partial}{\partial x}\{Var(\hat{\beta}_0) + x^2 Var(\hat{\beta}_1) + 2x\,Cov(\hat{\beta}_0, \hat{\beta}_1)\} = 0 \quad (1)$$

$$= 2x\,Var(\hat{\beta}_1) + 2Cov(\hat{\beta}_0, \hat{\beta}_1) = 0$$

$$\therefore x0 = -\frac{Cov(\hat{\beta}_0, \hat{\beta}_1)}{Var(\hat{\beta}_1)} = -\frac{-0.02668}{0.000579} = 46.08 \quad (2)$$

●(2)式の性質を少し詳しく確認してみよう。Page34-35②式より分散推定値は③式で推定される。

$$\hat{\mathrm{var}}\left[\hat{\beta}\right]=\hat{\mathrm{I}}^{-1}(\hat{\beta})=(\mathrm{X'VX})^{-1} \quad ②$$

$$\mathrm{X'VX}=\begin{bmatrix}1 & 1 & \cdots 1\\ x_1 & x_2 & \cdots x_n\end{bmatrix}\begin{bmatrix}\hat{\pi}_1(1-\hat{\pi}_1) & & & \\ & \hat{\pi}_2(1-\hat{\pi}_2) & & \\ & & \ddots & \\ & & & \hat{\pi}_n(1-\hat{\pi}_n)\end{bmatrix}\begin{bmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_n\end{bmatrix}$$

$$=\begin{bmatrix}\sum\hat{\pi}_i(1-\hat{\pi}_i) & \sum x_i\hat{\pi}_i(1-\hat{\pi}_i)\\ \sum x_i\hat{\pi}_i(1-\hat{\pi}_i) & \sum x_i^{\,2}\hat{\pi}_i(1-\hat{\pi}_i)\end{bmatrix}$$

$$\hat{\mathrm{var}}\left[\hat{\beta}\right]=(\mathrm{X'VX})^{-1}=\frac{1}{\alpha}\begin{bmatrix}\sum x_i^{\,2}\hat{\pi}_i(1-\hat{\pi}_i) & \sum x_i\hat{\pi}_i(1-\hat{\pi}_i)\\ \sum x_i\hat{\pi}_i(1-\hat{\pi}_i) & \sum\hat{\pi}_i(1-\hat{\pi}_i)\end{bmatrix} \quad ③$$

● ③式の結果を(2)式に代入すれば、(3)式が得られることになる。

$$x_0=-\frac{Cov(\hat{\beta}_0,\hat{\beta}_1)}{Var(\hat{\beta}_1)}=-\frac{-\frac{1}{\alpha}\sum x_i\ \hat{\pi}_i(1-\hat{\pi}_i)}{\frac{1}{\alpha}\sum\hat{\pi}_i(1-\hat{\pi}_i)}=\frac{-\sum x_i\ \hat{\pi}_i(1-\hat{\pi}_i)}{\sum\hat{\pi}_i(1-\hat{\pi}_i)}=-\frac{-\sum x_i\ \hat{\pi}_i(1-\hat{\pi}_i)}{n\bullet mean(\hat{\pi}_\bullet(1-\hat{\pi}_\bullet))} \quad (3)$$

今、$\hat{\pi}_1=\hat{\pi}_2=\cdots=\hat{\pi}_i=\cdots=\hat{\pi}_n=\hat{\pi}_\bullet$と$\hat{\pi}_i$がすべて等しければ上式は

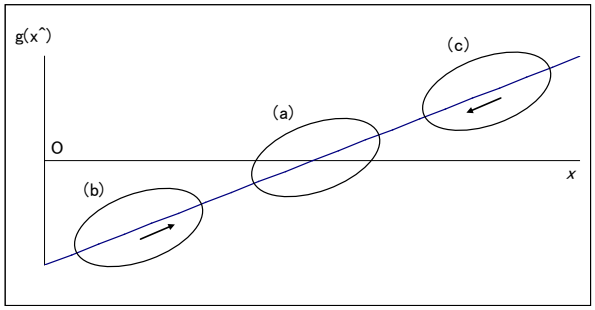
$x_0=-\dfrac{-\sum x_i}{n}=\bar{x}$と平均が一番分散が小さいことがわかる。

それ以外の場合は、xi の$\hat{\pi}_i(1-\hat{\pi}_i)$を重みとする、x の重み付け平均 x0 が最も分散の小さな x の値であることがわかる。

まとめ

● $\hat{\pi}_1=\hat{\pi}_2=\cdots=\hat{\pi}_i=\cdots=\hat{\pi}_n=\hat{\pi}_\bullet$と$\hat{\pi}_i$がすべて等しければ、x の平均が g(x)推定値の最小分散の位置であることがわかる。

●それ以外の場合は、xi の$\hat{\pi}_i(1-\hat{\pi}_i)$を重みとする x の重み付け平均 x0 が、最も分散の小さな x の値であることがわかる。つまり、$\hat{\pi}_i$=0.5 が最も重みが重く、1 に近づくほど重みが小さくなる x の重み付け平均 x0 が g(x)推定値の最小分散の位置であることがわかる。



図のような、データの頻度が x の範囲でほぼ等しい場合は、g(x)推定値の最小分散の位置は、x の平均よりも $\hat{g}(x)=0(\hat{\pi}=0.5)$ の方向に *x0* が移動していく（by　西野）。

また、例えば、x の分布が右に裾を引く対数正規のような場合は、仮に x の中央値よりも右に $\hat{\pi}_i$ =0.5 の点があれば、x の平均よりも *x0* は右に移動する。逆に、左に裾を引く分布で x の中央値よりも左に $\hat{\pi}_i$ =0.5 の点があれば、x の平均よりも *x0* は左に移動する。

以上が、単変量 Logistic model でロジット推定量 $\hat{g}(x)$ の信頼区間を最小にする変数の値 *x0* の性質である。