

カウント、フラクショナル、およびその他の 非負反応モデル

Instructor: Daisuke Nagakura
E-mail: nagakura@keio.jp

カウント、フラクショナル、およびその他の非負反応モデル

■ カウントデータとカウント変数

カウントデータとはカウント変数(計数変数ともいう)からなるデータのことである。カウント変数とは非負の整数値を取る変数のことである。

例えば

- ある人が所与の1年間に逮捕された回数、
- ある人が一日に吸ったたばこの本数、
- ある会社が所与の1年の間に申請した特許数、
などである。

カウント、フラクショナル、およびその他の非負反応モデル

カウントデータには2つの種類がある。

一つは、取りうる値に(自然なあるいは既知の)**上限を持たない**ものの、もう一つは**上限を持つもの**である。

前頁の例は全て前者の種類になる。

後者の例としては、例えば、ある家族における高校卒業した子供の数などがあげられる。この場合、上限はその家族の子供の数となる。

以後は、主に前者の種類のカウントデータを考える(後者も後ほど扱う)。

カウント、フラクショナル、およびその他の非負反応モデル

■ ポアソン分布

確率変数 y は非負の整数値を取り、さらに上限はないとする。この時、その確率関数が以下で与えられる時、この確率変数 y は**ポアソン分布**に従うという:

$$p(y) = \frac{\exp(-\lambda)\lambda^y}{y!}, \quad y = 0, 1, 2, \dots$$

ここで、 $\lambda > 0$ は分布のパラメーターである。この分布の期待値と分散は以下のように与えられる。

$$E(y) = \lambda, \quad \text{および} \quad \text{var}(y) = \lambda$$

これは、この分布においては**分散平均比** $\text{var}(y)/E(y)$ が1であることを意味する。以後、 y がパラメーター λ のポアソン分布に従うことを $y \sim \text{Po}(\lambda)$ のように書くこととする。

カウント、フラクショナル、およびその他の非負反応モデル

■ ポアソン回帰モデル

ポアソン回帰モデル は被説明変数 y_i を以下のように定式化する。

$$y_i \mid \mathbf{x}_i \sim \text{Po}(\lambda_i), \quad \lambda_i = m(\mathbf{x}_i, \boldsymbol{\beta})$$

ここで $m(\mathbf{x}_i, \boldsymbol{\beta}) > 0$ は説明変数ベクトル $\mathbf{x}_i$ のスカラーの関数である。この時、 y_i の $\mathbf{x}_i$ で条件付けした、条件付き確率関数は以下で与えられる。

$$p(y_i \mid \mathbf{x}_i) = \frac{\exp(-\lambda_i) \lambda_i^{y_i}}{y_i!}, \quad \lambda_i = m(\mathbf{x}_i, \boldsymbol{\beta})$$

この時、 $E(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i, \boldsymbol{\beta})$ であるので、この関数 $m(.,.)$ は(条件付き) **期待値関数** と呼ばれる。

カウント、フラクショナル、およびその他の非負反応モデル

■ 期待値関数の定式化

最もよく用いられる期待値関数の定式化は

$$m(\mathbf{x}_i, \boldsymbol{\beta}) = \exp(\mathbf{x}_i' \boldsymbol{\beta})$$

である。期待値関数がこのように定式化された時、 $\boldsymbol{\beta}$ の値は以下のように解釈できる。まず $m(\cdot)$ を x_{ik} で偏微分して

$$\partial m(\mathbf{x}_i, \boldsymbol{\beta}) / \partial x_{ik} = \exp(\mathbf{x}_i' \boldsymbol{\beta}) \beta_k,$$

を得る。よって

$$\beta_k = [\partial m(\mathbf{x}_i, \boldsymbol{\beta}) / \partial x_{ik}] / \exp(\mathbf{x}_i' \boldsymbol{\beta}) = \partial \log[m(\mathbf{x}_i, \boldsymbol{\beta})] / \partial x_{ik}.$$

となる。これは、 x_{ik} がわずかに変化した時、例えば Δx_{ik} だけ変化した時に、 y_i の期待値は約 $100 \times \beta_k \Delta x_{ik} \%$ 変化することを意味している。

カウント、フラクショナル、およびその他の非負反応モデル

■ 最尤推定

ポアソン回帰モデルの最尤推定値は以下のポアソンモデルの対数尤度関数を最大化することによって得られる(以下では推定に影響を与えない定数項を除いている)。

$$L_N^{(Po)}(\boldsymbol{\beta}) = \sum_{i=1}^N \ell_i^{(Po)}(\boldsymbol{\beta}), \text{ ここで } \ell_i^{(Po)}(\boldsymbol{\beta}) = y_i \log[m(\mathbf{x}_i, \boldsymbol{\beta})] - m(\mathbf{x}_i, \boldsymbol{\beta}).$$

この時、(最大化のための)1階の条件は

$$\mathbf{S}_N^{(Po)}(\boldsymbol{\beta}) = \sum_{i=1}^N \mathbf{s}_i^{(Po)}(\boldsymbol{\beta}) = \mathbf{0} \text{ ここで } \mathbf{s}_i^{(Po)}(\boldsymbol{\beta}) = \frac{\partial m(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \left[\frac{y_i}{m(\mathbf{x}_i, \boldsymbol{\beta})} - 1 \right]$$

で与えられる。例えば $m(\mathbf{x}_i, \boldsymbol{\beta}) = \exp(\mathbf{x}_i' \boldsymbol{\beta})$ である時には

$$\mathbf{s}_i^{(Po)}(\boldsymbol{\beta}) = \exp(\mathbf{x}_i' \boldsymbol{\beta}) [y_i \exp(-\mathbf{x}_i' \boldsymbol{\beta}) - 1]$$

となる。通常、この計算はコンピューターを用いて数值的に行う。

カウント、フラクショナル、およびその他の非負反応モデル

■ ポアソン回帰モデルの最尤推定量の漸近分布

最尤推定の一般的な結果を適用すると、 y_i が実際に $Po(\lambda_i)$ に従っている時には β の最尤推定量 $\hat{\beta}$ の漸近分布は

$$\sqrt{N}(\hat{\beta} - \beta_0) \sim_A N(\mathbf{0}, \mathbf{A}_0^{-1})$$

で与えられる。ここで β_0 は β の真の値、行列 $\mathbf{A}_0$ は

$$\mathbf{A}_0 = -E \left[\mathbf{H}_i^{(Po)}(\beta) \right] = -E \left\{ E \left[\mathbf{H}_i^{(Po)}(\beta) \mid \mathbf{x}_i \right] \right\},$$

で与えられる。ここで $\mathbf{H}_i^{(Po)}(\beta) = \frac{\partial^2 \ell_i^{(Po)}}{\partial \beta \partial \beta'}$ である。情報行列等式より

$$\mathbf{A}_0 = \mathbf{B}_0 = E \left[\mathbf{s}_i^{(Po)}(\beta) \mathbf{s}_i^{(Po)'}(\beta) \right] = E \left\{ E \left[\mathbf{s}_i^{(Po)}(\beta) \mathbf{s}_i^{(Po)'}(\beta) \mid \mathbf{x}_i \right] \right\},$$

が得られることに注意しよう。ここで $\mathbf{s}_i^{(Po)}(\beta) = \frac{\partial \ell_i^{(Po)}}{\partial \beta}$ である。

カウント、フラクショナル、およびその他の非負反応モデル

Exercise 1. Suppose that $y_i \sim Po(m(\mathbf{x}_i, \boldsymbol{\beta}))$. Then

- (1) Show that $E[\mathbf{s}_i^{(Po)}(\boldsymbol{\beta}) \mid \mathbf{x}_i] = \mathbf{0}$.
- (2) Calculate the Hessian matrix $\mathbf{H}_i^{(Po)}(\boldsymbol{\beta}) = \partial \mathbf{s}_i^{(Po)}(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}'$.
- (3) Calculate the $-E[\mathbf{H}_i^{(Po)}(\boldsymbol{\beta}) \mid \mathbf{x}_i]$.
- (4) Conform the information matrix equality holds when $y_i \sim Po(m(\mathbf{x}_i, \boldsymbol{\beta}))$.

カウント、フラクショナル、およびその他の非負反応モデル

■ 疑似最尤推定 (Quasi maximum likelihood estimation; QML)

先ほどは y_i が実際にポアソン回帰モデルに従っていると仮定した。しかしながら、この仮定が正しいかは実際にはわからない。もし y_i が実際にはポアソン回帰モデルに従っていない場合でも、ポアソン回帰モデルの最尤推定量は、**期待値関数の定式化が正しい限りは**、実際の分布が何であれ（さらに、 y_i の分布が非負の連続型の確率変数の時でさえ）、パラメーター β の**一致推定量**となる。

y_i が必ずしもポアソン回帰モデルに従っていないと想定する場合、ポアソン回帰モデルの最尤推定量は**ポアソンQML推定量**と呼ばれる。

ただし、この場合は、漸近分散は先ほどと異なるため、推定にはサンドイッチ型の推定量を用いる必要がある。

カウント、フラクショナル、およびその他の非負反応モデル

■ ポアソンQML推定量の漸近分布

疑似最尤推定量の一般的な結果を適用することにより、ポアソンQML推定量に対して、以下の結果が得られる。

$$\sqrt{N}(\hat{\beta} - \beta_0) \sim N(\mathbf{0}, \mathbf{A}_0^{-1} \mathbf{B}_0 \mathbf{A}_0^{-1}),$$

ここで、 $\mathbf{A}_0$, $\mathbf{B}_0$ は先ほどと同様に定義される。ただし、そこでの期待値は y_i の真の分布について取られる。これは必ずしもポアソン分布ではなく、また実際にはわからないことに注意しよう。

カウント、フラクショナル、およびその他の非負反応モデル

■ ポアソン疑似最尤推定量の効率性

ポアソンQML推定量は、 y_i が実際にポアソン回帰モデルに従っていれば最尤推定量であるので、その意味で効率的である。

しかしながら、ポアソンQML推定量は、実際の y_i がポアソン回帰モデルに従っていない場合でも、以下の条件のもとで、ある限定された推定量のクラスの中で、効率的となる。この仮定は **ポアソンGLM分散の仮定** と呼ばれる(ここでGLMは generalized linear model; 一般化線形モデルの略)

$$\text{var}(y_i | \mathbf{x}_i) = \sigma^2 E(y_i | \mathbf{x}_i),$$

ここで、定数 $\sigma^2 > 0$ は分散 - 期待値 比である。

この条件の下で、ポアソンQML推定量は y_i が線形指数分布族に従っている場合は、最も効率的な推定量になる。

カウント、フラクショナル、およびその他の非負反応モデル

■ ポアソンQML推定量の漸近分散

ポアソンGLM分散の仮定が成り立っている時には、ポアソンQML推定量の漸近分散は

$$\text{Avar}(\sqrt{N}(\hat{\beta} - \beta_0)) = \sigma^2 \mathbf{A}_0^{-1},$$

となる。ここで、 $\mathbf{A}_0$ 先ほどと同じ定義であり、 σ^2 はポアソンGLM分散の仮定の中の σ^2 である。定義より $\sigma^2 = \text{var}(y_i | \mathbf{x}_i) / E(y_i | \mathbf{x}_i)$ であるので、この定数は

$$\hat{\sigma}^2 = \sum_{i=1}^N \hat{u}_i^2 / m(\mathbf{x}_i, \hat{\beta})$$

によって一致推定できる。ここで $\hat{u}_i = y_i - m(\mathbf{x}_i, \hat{\beta})$ である。よって上記の漸近分散の一致推定量として

$$\text{Avar}(\sqrt{N}(\hat{\beta} - \beta_0)) = \hat{\sigma}^2 \hat{\mathbf{A}}^{-1}, \text{ ここで } \hat{\mathbf{A}} = N^{-1} \sum_{i=1}^N \mathbf{H}_i^{(Po)}(\hat{\beta})$$

を得る。この推定量は、しかしながら、ポアソンGLM分散の仮定が満たされていない時には一致推定量ではないことに注意しよう。

カウント、フラクショナル、およびその他の非負反応モデル

Exercise 2. Suppose that $y_i \sim Po(m(\mathbf{x}_i, \boldsymbol{\beta}))$. Then
Confirm that the asymptotic variance of the **Poisson QML estimator** is given as

$$\text{Avar}(\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)) = \sigma^2 \mathbf{A}_0^{-1},$$

when the Poisson GLM assumption is satisfied.

カウント、フラクショナル、およびその他の非負反応モデル

■ 二項回帰モデル

ここまでは、自然な上限がないカウント変数について考えてきた。ここでは、**既知の上限**があるカウントデータについて考える。ここではそのような変数に対して、**二項分布**をベースにしたモデルを考える。

カウント変数 y の取りうる値は $0, 1, \dots, n$ の $n+1$ 個であるとしよう。この場合、 y の上限は n である。 y の確率関数が

$$\Pr(y_i = k) = {}_k C_n p^k (1-p)^{n-k}, \quad {}_k C_n = \frac{n!}{k!(n-k)!}$$

で与えられる時、 y は二項分布に従うという。ここで n と p は分布のパラメーターである。

以降は y がパラメーター n と p の二項分布に従うことを $y \sim B(p, n)$ と書くことにする。

カウント、フラクショナル、およびその他の非負反応モデル

二項分布に従う確率変数 y は

$$y = x_1 + x_2 + \dots + x_n,$$

のように表すことができる。ここで、 $x_i, i=1, \dots, n$, は互いに独立な n 個のベルヌーイ確率変数であり、確率 p で $x_i = 1$ 、確率 $1 - p$ で $x_i = 0$ を取るとする。この結果より、

$$E(y) = np \quad \text{および} \quad \text{var}(y) = np(1 - p)$$

が簡単に求まる。

上記の表現は、二項分布が、 n 回の**ベルヌーイ試行**における**成功回数**の分布であることを意味している。ベルヌーイ試行とは、起きうる結果が、「成功」と「失敗」の2つしかなく、成功は確率 p 、失敗は確率 $1 - p$ で起こるような試行のことである。

カウント、フラクショナル、およびその他の非負反応モデル

二項回帰モデルは y_i を $y_i \sim B(p(\mathbf{x}_i, \boldsymbol{\beta}), n_i)$ と定式化したモデルである。ここで $p(\mathbf{x}_i, \boldsymbol{\beta})$ は 0と1の間の値を取るスカラー関数である。

ここで n_i は **個体 i 毎に異なっている**ことに注意しよう。また、 n_i は **既知**であるとする。よってモデルの未知パラメーターは $p(\mathbf{x}_i, \boldsymbol{\beta})$ に含まれる $\boldsymbol{\beta}$ のみとなる。

通常、関数 $p(.,.)$ は $p(\mathbf{x}_i, \boldsymbol{\beta}) = G(\mathbf{x}_i' \boldsymbol{\beta})$ と定式化される。ここで $G(.)$ は適当な連続型確率変数の分布関数である。

y_i の $\mathbf{x}_i$ で条件付けした条件付き期待値と分散は

$$E(y_i | \mathbf{x}_i, n_i) = n_i p(\mathbf{x}_i, \boldsymbol{\beta}) \text{ および}$$

$$\text{var}(y_i | \mathbf{x}_i, n_i) = n_i p(\mathbf{x}_i, \boldsymbol{\beta})(1 - p(\mathbf{x}_i, \boldsymbol{\beta}))$$

で与えられる。

カウント、フラクショナル、およびその他の非負反応モデル

■ 二項回帰モデルの最尤推定

二項回帰モデルの対数尤度関数は(推定に関係ない部分を除く)

$$L_N^{(B)}(\boldsymbol{\beta}) = \sum_{i=1}^N \ell_i^{(B)}(\boldsymbol{\beta})$$

で与えられる。ここで

$$\ell_i^{(B)}(\boldsymbol{\beta}) = y_i \log p(\mathbf{x}_i, \boldsymbol{\beta}) + (n_i - y_i) \log(1 - p(\mathbf{x}_i, \boldsymbol{\beta}))$$

である。通常のML推定の理論より、最尤推定量 $\hat{\boldsymbol{\beta}}$ は漸近的に正規分布に従うことを示すことができる。

カウント、フラクショナル、およびその他の非負反応モデル

■ 疑似最尤推定量としての二項回帰モデル

ポアソン回帰モデルの時と同様、二項回帰モデルの最尤推定量も、例えば y_i が**二項分布に従っていない**場合でも、期待値関数が以下の条件を満たせば β の**一致推定量**となる。

$$E(y_i | \mathbf{x}_i, n_i) = n_i p(\mathbf{x}_i, \beta)$$

つまり期待値関数が $n_i p(\mathbf{x}_i, \beta)$ で正しく定式化されているという条件である。特に、 $p(.,.)$ が正しく定式化されていることに注意しよう。

二項回帰モデルからの疑似最尤指定量の漸近分散も、通常のサンドイッチフォーム推定量により一致推定できる。

カウント、フラクショナル、およびその他の非負反応モデル

Exercise 3.

- (1) Derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ in the binomial regression model when y_i actually follows $B(p(\mathbf{x}_i, \boldsymbol{\beta}_0), n_i)$.
- (2) Derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ in the binomial regression model when y_i actually does not follow $B(p(\mathbf{x}_i, \boldsymbol{\beta}_0), n_i)$, but the mean function is correctly specified as $n_i p(\mathbf{x}_i, \boldsymbol{\beta})$,

カウント、フラクショナル、およびその他の非負反応モデル

■ 非負値を取る連続型の被説明変数

ここでは、 y_i は**自明な上限がない非負の連続型の変数**である場合を考える。

この場合でも、ポアソン回帰(つまりポアソンQML)推定量は、 $E(y_i | \mathbf{x}_i) = m(\mathbf{x}_i, \beta)$ が正しく定式化されていれば、 β の一致推定量である(ポアソン回帰はカウント変数用のモデルであるが)。

また、ポアソン回帰推定量は、ポアソンGLM分散の仮定が満たされれば、ある特定のクラスの推定量の中で最も効率的である。ただしこの仮定は非負の連続値を取る変数に対しては成り立つ場合は少ないと思われる。

カウント、フラクショナル、およびその他の非負反応モデル

■ 連続値を取る変数に対する慣例的な方法

変数 y_i は自明な上限のない正の連続値を取る変数とする。

この場合、慣例的な方法として、 y_i の対数を β の線形の関数としてモデル化するという方法がある。すなわち $\log y_i$ を

$$\log y_i = \mathbf{x}_i' \boldsymbol{\beta} + u_i ,$$

と定式化する。ここで u_i は期待値 0 の誤差項であり、 $\mathbf{x}_i$ と独立であるとする。

このモデルに対して、 β は通常の方法によって推定できる。

カウント、フラクショナル、およびその他の非負反応モデル

このモデルに対して、所与の $\mathbf{x}_i$ に対して、 y_i の期待値は

$$\begin{aligned} E(y_i | \mathbf{x}_i) &= E[\exp(\mathbf{x}_i' \boldsymbol{\beta} + u_i)] \\ &= \exp(\mathbf{x}_i' \boldsymbol{\beta}) E[\exp(u_i)] \\ &= \exp(\mathbf{x}_i' \boldsymbol{\beta}) \alpha, \end{aligned}$$

で与えられる。ここで $\alpha = E[\exp(u_i)]$ である。この α は

$$\hat{\alpha} = N^{-1} \sum_{i=1}^N \exp(\hat{u}_i),$$

によって推定できる。ここで $\hat{u}_i$ はこの線形回帰モデルの残差である。また y_i の $\mathbf{x}_i$ で条件付けした分散は

$$\text{var}(y_i | \mathbf{x}_i) = \kappa [\exp(\mathbf{x}_i' \boldsymbol{\beta})]^2,$$

によって与えられる。ここで $\kappa = \text{var}[\exp(u_i)]$ である。

カウント、フラクショナル、およびその他の非負反応モデル

注意すべき点として、ここでは u_i は $\mathbf{x}_i$ と独立と仮定しているが、この仮定は必ずしも成り立たない可能性がある。

このような場合、例えば $u_i | \mathbf{x}_i \sim N(0, \exp(\mathbf{x}_i' \boldsymbol{\gamma}))$ である時には、この場合 $\exp(u_i)$ は期待値 $\exp(\mathbf{x}_i' \boldsymbol{\gamma}/2)$ の対数正規分布に従うので

$$\begin{aligned} E(y_i | \mathbf{x}_i) &= \exp(\mathbf{x}_i' \boldsymbol{\beta}) E[\exp(u_i) | \mathbf{x}_i] \\ &= \exp(\mathbf{x}_i' \boldsymbol{\beta} + \exp(\mathbf{x}_i' \boldsymbol{\gamma})/2) \end{aligned}$$

となり、条件付き期待値は先ほどと異なってくる。

カウント、フラクショナル、およびその他の非負反応モデル

■ ガンマ回帰モデル

より直接的な方法として、 $E(y_i | \mathbf{x}_i)$ および y_i の分布を直接定式化するという方法が考えられる。

ガンマ回帰モデルは、説明変数 $\mathbf{x}_i$ で条件付けした時に、 y_i は**ガンマ分布**に従うとして y_i をモデル化する。

ある正値をとる確率変数 y が下記のpdfを持つとき y はガンマン分布に従うという:

$$f(y) = \frac{1}{\Gamma(\alpha)\theta^\alpha} y^{\alpha-1} \exp(-y / \theta), \alpha > 0, \theta > 0$$

この時、 y の期待値と分散は $E(y) = \alpha\theta$ および $\text{var}(y) = \alpha\theta^2$ で与えられる。以後この分布を $G(\alpha, \theta)$ と表す。

カウント、フラクショナル、およびその他の非負反応モデル

特に、ガンマ回帰モデルでは、 y_i を

$$y_i \mid \mathbf{x}_i \sim G(1, m(\mathbf{x}_i, \beta))$$

と定式化する。この時、 $E(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i, \beta)$ および $\text{var}(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i, \beta)^2$ である。

$\alpha = 1$ の時、ガンマ分布は**指数分布**となるので、上記の定式化は $y_i \sim \text{Exp}(m(\mathbf{x}_i, \beta))$ と等しい。ここで $\text{Exp}(\theta)$ はパラメーター θ の指数分布を表す。よって上記のガンマ回帰モデルは**指数回帰モデル**とも呼ばれる。

ガンマ回帰モデルの対数尤度関数は(定数項を除いて)以下で与えられる。

$$L_N^{(G)}(\boldsymbol{\beta}) = \sum_{i=1}^N \ell_i^{(G)}(\boldsymbol{\beta}), \quad \ell_i^{(G)}(\boldsymbol{\beta}) = -y_i / m(\mathbf{x}_i, \boldsymbol{\beta}) - \log[m(\mathbf{x}_i, \boldsymbol{\beta})]$$

カウント、フラクショナル、およびその他の非負反応モデル

■ Gamma regression as a QML estimation

Gamma regression is actually a consistent estimator for β even when y_i does not follow the specified gamma distribution as long as the conditional mean $E(y_i | \mathbf{x}_i) = m(\mathbf{x}_i, \beta)$ is correctly specified.

In this context, the gamma regression is called the **gamma QML estimation**.

カウント、フラクショナル、およびその他の非負反応モデル

Furthermore, if the variance of the population distribution of y_i given $\mathbf{x}_i$ satisfies the so-called **gamma GLM variance assumption**, i.e.,

$$\text{var}(y_i | \mathbf{x}_i) = \sigma^2 [E(y_i | \mathbf{x}_i)]^2, \quad \sigma^2 > 0,$$

then the gamma QML estimation is efficient in a class of estimators that correctly specifies only $E(y_i | \mathbf{x}_i)$.

In the above equation, $\sigma = \text{var}(y_i | \mathbf{x}_i)^{1/2} / E(y_i | \mathbf{x}_i)$, which is so-called the (conditional) **coefficient of variation**: it is the ratio of the (conditional) standard deviation to its (conditional) mean.

カウント、フラクショナル、およびその他の非負反応モデル

Exercise 4.

- (1) Suppose $x \sim G(\alpha, 1)$. Define a new random variable y as $y = \theta x$. Show that $y \sim G(\alpha, \theta)$. Thus θ is called scale parameter.
- (2) Derive the score and Hessian matrix of the Gamma QML log likelihood estimation.

カウント、フラクショナル、およびその他の非負反応モデル

■ フラクショナル反応モデル

最後に、変数 y_i として、取りうる値の範囲が $[0, 1]$ である(ただし y_i は二値変数でない)モデルを考える。

上記において変数 y_i の取りうる範囲が $(0, 1)$ の場合(つまり0と1を含まない)、慣例的なやり方は y_i を地域が $(-\infty, \infty)$ となるように変換して、変換後の変数に対して線形回帰モデルを当てはめるというものである。このような変換として対数オッズ変換

$$w_i = \log[y_i / (1 - y_i)],$$

が良く用いられる。この w_i に対して線形回帰モデル

$$w_i = \mathbf{x}_i' \boldsymbol{\beta} + u_i$$

を当てはめるのである。

カウント、フラクショナル、およびその他の非負反応モデル

この2つの式より $\log[y_i / (1 - y_i)] = \mathbf{x}_i' \boldsymbol{\beta} + u_i$ が得られるので、この式を y_i について解くことにより

$$y_i = \exp(\mathbf{x}_i' \boldsymbol{\beta} + u_i) / [1 + \exp(\mathbf{x}_i' \boldsymbol{\beta} + u_i)]$$

をえる。これより $E(y_i | \mathbf{x}_i = \mathbf{x})$ は

$$N^{-1} \sum_{k=1}^N \frac{\exp(\mathbf{x}' \hat{\boldsymbol{\beta}} + \hat{u}_k)}{1 + \exp(\mathbf{x}' \hat{\boldsymbol{\beta}} + \hat{u}_k)},$$

によって推定することができる。ここで $\hat{\boldsymbol{\beta}}$ は $w_i = \mathbf{x}_i' \boldsymbol{\beta} + u_i$ の最小二乗推定値であり、 $\hat{u}_i = w_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}$ はこの回帰モデルの残差である。

対数オッズ変換の他にも、標準正規分布の分布関数 $\Phi(\cdot)$ の逆関数もよく用いられる。

カウント、フラクショナル、およびその他の非負反応モデル

■ フラクショナル回帰ロジット(プロビット)回帰

このような慣例的な方法の代わりに、このような $[0, 1]$ の範囲を取るようなデータ y_i に対して、いわゆる**フラクショナルロジット(プロビット)回帰**と呼ばれる方法を用いることもできる(これは y_i が連続でなくてもよい)。

フラクショナル回帰モデルでは $E(y_i | \mathbf{x}_i)$ を

$$E(y_i | \mathbf{x}_i) = \exp(\mathbf{x}_i' \boldsymbol{\beta}) / [1 + \exp(\mathbf{x}_i' \boldsymbol{\beta})] \text{ (フラクショナルロジット)},$$

または

$$E(y_i | \mathbf{x}_i) = \Phi(\mathbf{x}_i' \boldsymbol{\beta}) \text{ (フラクショナルプロビット)},$$

のように定式化する。

カウント、フラクショナル、およびその他の非負反応モデル

■ フラクショナル回帰モデルの推定

フラクショナル回帰モデルにおいて β の推定には以下のような疑似対数尤度関数を用いる。

$$L_N^{(F)}(\beta) = \sum_{i=1}^N \ell_i^{(F)}(\beta),$$

$$\ell_i^{(F)}(\beta) = (1 - y_i) \log[1 - G(\mathbf{x}_i' \beta)] + y_i \log[G(\mathbf{x}_i' \beta)],$$

ここで $G(\mathbf{x}_i' \beta) = E(y_i | \mathbf{x}_i)$ である。

慣例的な方法では、0 と 1 の値は取らないという制約があったのに対して、フラクショナル回帰モデルでは y_i は $[0, 1]$ 上のどんな数 (0 と 1 を含む) を取っても大丈夫というアドバンテージがある。

カウント、フラクショナル、およびその他の非負反応モデル

フラクショナル回帰推定量は、疑似最尤推定量であるが、**(条件付き)期待値関数が正しく定式化されていれば** β の一致推定量であり、以下の**ベルヌーイGLM分散の仮定**が成り立つ時には、(条件付き)期待値関数が正しく定式化されているモデルに対して、最も効率的な推定量である。

(ベルヌーイGLM分散の仮定)

$$\text{var}(y_i | \mathbf{x}_i) = \sigma^2 E(y_i | \mathbf{x}_i)[1 - E(y_i | \mathbf{x}_i)].$$

フラクショナル回帰推定は、**ベルヌーイQML推定**とも呼ばれる(なぜなら、この疑似対数尤度は y_i にベルヌーイ分布を仮定したものである。ただし、フラクショナル回帰推定においては y_i は $y_i \in [0, 1]$ であれば、0 と 1 以外の値もとってもよい)。

カウント、フラクショナル、およびその他の非負反応モデル

Exercise 5.

(1) Show that when $y_i = s_i/n_i$, where $s_i \sim B(p(\mathbf{x}_i, \boldsymbol{\beta}), n_i)$, the Bernoulli GLM assumption holds for this y_i .