統計解析

古谷知之

授業概要

* 履修者の状況に応じて変更される場合がありますが、全体としては以下のような授業構成となります。
* 講義の中でR演習を行うこともあります。

第1回　ガイダンス・単回帰分析
第2回　重回帰分析(1)
第3回　重回帰分析(2)
第4回　一般化線形回帰モデル(1)
第5回　一般化線形回帰モデル(2)
第6回　一般化線形回帰モデル(3)
第7回　一般化線形回帰モデル(4)
第8回　一般化線形回帰モデル(5)
第9回　一般化線形回帰モデル(6)
第10回　一般化線形混合モデル
第11回　状態空間モデル
第12回　R演習(1)
第13回　R演習(2)
第14回　R演習(3)

統計モデルの種類

	主な推定方法	データ分布	回帰係数
線形回帰モデル （単回帰・重回帰など）	最小二乗法	正規分布	一変数に一つ
一般化線形モデル	最尤推定法	正規分布以外の分布も可能	一変数に一つ
一般化線形混合モデル	最尤推定法	正規分布以外の分布も可能	変数の個体差に応じて推定可能
階層ベイズモデル	ベイズ推定	正規分布以外の分布も可能	変数の個体差に応じて推定可能

本授業で扱う統計モデル

- 線形回帰モデル
  - 単回帰モデル、重回帰モデル
- 一般化線形回帰モデル
  - 離散：ポアソン回帰モデル、二項反応モデル（ロジスティック回帰モデル、プロビット回帰モデル、補対数対数モデル）、負の二項分布モデル、ゼロ過剰ポアソン回帰モデル、ゼロ過剰負の二項分布モデル
  - 連続：ガンマ回帰モデル、ベータ回帰モデル、指数-ガウス回帰モデル
  - スパース：Lasso回帰モデル、Ridge回帰モデル
- 一般化線形混合モデル
  - マルチレベルモデル
- 状態空間モデル

代表的な一般化線形回帰モデル

- 被説明変数が離散変数
  - 0or1の２値：(二項)ロジスティック回帰モデル、(二項)プロビット回帰モデル、補対数対数モデル
  - ０以上の整数
    - 発生頻度が少ない：ポアソン回帰モデル、負の二項分布モデル
    - 発生頻度０が非常に多い：Hurdleモデル、ゼロ過剰モデル
- 被説明変数が連続変数
  - [0, 1]の確率値：ベータ回帰モデル
  - ０より大きい値：ガンマ回帰モデル、指数-ガウス回帰モデル
- 被説明変数がスパース
  - Lasso回帰モデル、Ridge回帰モデル

ゼロの発生頻度が多いモデル

- 非負の整数となる被説明変数のうち、０の値を取るケースが非常に多い場合、カウントデータモデルを推定する際には０か０でない（０より大きい）かを把握することが重要となる

- このような場合、以下の２種類のモデルが用いられる

- Hurdleモデル（ゼロ切断モデル）
- ゼロ過剰モデル（ゼロ強調モデル/zero-inflated model）

Hurdleモデルとゼロ過剰モデル

- Hurdleモデル
  - 0か1以上かを判別するモデル＋発生頻度モデル
  - 1以上の値となる（0を超える）までハードルがある
  - 0であることはきちんと観測されている
- ゼロ過剰モデル
  - 0でない可能性を判別するモデル＋発生頻度モデル
  - 本来は0でないかもしれないが、0と観測されたデータが多い
  - 0であることは偶然観測された可能性がある

- 発生頻度モデルにはポアソン回帰モデル、負の二項分布モデルなどが用いられる

Hurdleモデルとゼロ過剰モデル

- Hurdleモデル
  - 0か1以上かを判別する二項分布モデルと切断モデルとを分けて用いる
  - 係数が正なら0が減り、1以上になりやすい

- ゼロ過剰モデル
  - 0か1以上かを判別する二項分布モデルとポアソン回帰モデルなどとの混合モデル
  - 係数が正なら0が増え、偶然観測された0である可能性が高まる

Hurdleモデル（ゼロ切断モデル）

- 確率変数Yが$y=0$となる確率を$\pi\ (0<\pi<1)$、１以上の整数値$(y=1,2,\ldots)$をとる確率を$1-\pi$とすると、hurdleモデルではYが$y=0$となるかどうかの二項過程を次式のように表す

$$P(Y=y)=\begin{cases}\pi, & y=0\\ 1-\pi, & y=1,2,3,\ldots\end{cases}$$

- $y=0$となるリンク関数を$g_1(\pi_i)$、$y=1,2,3,\ldots$となるリンク関数を$g_2(\lambda_i)$とすると、ゼロ切断モデルは以下のように表せる

$$P(Y_i=y_i)=\begin{cases}g_1(\pi_i), & y=0\\ \big(1-g_1(\pi_i)\big)g_2(\lambda_i), & y=1,2,3,\ldots\end{cases}$$

- $g_1(\pi_i)$と$g_2(\lambda_i)$はそれぞれ線形予測子で表す

ゼロ過剰モデル（ゼロ強調モデル）

- ゼロ過剰モデルでは観測値$y=0$が偶然である（実際には$y>0$かもしれない）可能性を排除しないため、$Y=y$となる関数$f(y)$を用いて、次式のように表す

$$P(Y=y)=\begin{cases}\pi+(1-\pi)f(y), & y=0\\ (1-\pi)f(y), & y=1,2,3,\ldots\end{cases}$$

- $y=0$となるリンク関数を$g_1(\pi_i)$、$y=1,2,3,\ldots$となるリンク関数を$g_2(\lambda_i)$とすると、ゼロ切断モデルは以下のように表せる

$$P(Y_i=y_i)=\begin{cases}g_1(\pi_i)+\big(1-g_1(\pi_i)\big)g_2(\lambda_i), & y=0\\ \big(1-g_1(\pi_i)\big)g_2(\lambda_i), & y=1,2,3,\ldots\end{cases}$$

- $g_1(\pi_i)$と$g_2(\lambda_i)$はそれぞれ線形予測子で表す

（二項）ロジスティック回帰モデル

- ベルヌーイ試行に従う被説明変数(0または1の二値)を説明するロジスティック回帰モデルは、例えば以下のように表せる

$$y_i \sim Binom(n,p) \approx \pi_i^r \cdot (1-\pi_i^r)^{n-r}$$

- リンク関数

$$g\big(E(y_i)\big) = g(\pi_i) = \frac{1}{1+\exp\big(-(\eta_i)\big)}$$

$$\mathrm{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$$

- 線形予測子

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = X_i^T \boldsymbol{\beta}$$

ポアソン回帰モデル

- 被説明変数がポアソン分布に従うとするポアソン回帰モデルは、例えば次式のように表される。ここでy_iは被説明変数、X_iは説明変数、$\boldsymbol{\beta}$は未知パラメータである

$$y_i = Po(\lambda_i)$$

$$f(y_i|\lambda_i) = \frac{{\lambda_i}^{y_i} \cdot \exp(-\lambda_i)}{y_i!}$$

$$g\big(E(y_i)\big) = g(\lambda_i) = \exp(\eta_i)$$

- このときlogリンク(対数リンク)$\log(\lambda_i)$と線形予測子$\eta_i = X_i\boldsymbol{\beta}$との関係は、次式のように対数リンク関数で表せる

$$\log(\lambda_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = X_i^T \boldsymbol{\beta}$$

ゼロ切断ポアソン分布

- ポアソン分布は

$$f(y|\lambda) = \frac{\lambda^y \cdot \exp(-\lambda)}{y!}$$

- $y > 0$のとき$P(Y = y|Y > 0)$のprobability mass function (確率質量関数)は以下のようになる

$$P(Y = y|Y > 0) = \frac{f(y|\lambda)}{1 - f(0|\lambda)} = \frac{\lambda^y \cdot \exp(-\lambda)}{\big(1 - \exp(-\lambda)\big)y!} = \frac{\lambda^y}{(\exp(-\lambda) - 1)y!}$$

ゼロ切断ポアソン分布

- ゼロで切断されたポアソン分布は次式のように表せる

$$P(Y = y) = \begin{cases} \dfrac{\lambda^y}{(\exp(-\lambda) - 1)y!}, y = 1, 2, 3, \ldots \\ 0, \qquad\qquad otherwise \end{cases}$$

ゼロ切断ポアソン分布

- ゼロ切断ポアソン分布の平均$E(Y)$と分散$Var(Y)$は以下のとおり

$$E(Y) = \frac{\lambda}{1 - \exp(-\lambda)} = \frac{\lambda \cdot \exp(\lambda)}{(\exp(-\lambda) - 1)}$$

$$Var(Y) = \frac{\lambda}{1 - \exp(-\lambda)}\left[(1 + \lambda) - \frac{\lambda}{1 - \exp(-\lambda)}\right]$$
$$= \frac{\lambda \cdot \exp(\lambda)}{(\exp(-\lambda) - 1)}\left[1 - \frac{\lambda \cdot \exp(\lambda)}{(\exp(-\lambda) - 1)}\right]$$

Hurdleポアソン回帰モデル

- Hurdleポアソン回帰モデル（ゼロ切断ポアソン回帰モデル）は、次式のように表せる

$$P(Y_i = y_i) = \begin{cases} \pi_i, & y = 0 \\ (1 - \pi_i)\dfrac{{\lambda_i}^{y_i}}{(\exp(-\lambda_i) - 1)y_i!}, & y = 1, 2, 3, \ldots \end{cases}$$

- π_iはロジットリンクで、説明変数行列Zの要素z_{ij}と未知パラメータδ_jを用いて線形予測子は次式のように表せる

$$\mathrm{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = z_{i0}\delta_0 + z_{i1}\delta_1 + \cdots + z_{ik}\delta_k$$

- λ_iの対数リンクと線形予測子は次式のようになる

$$\log(\lambda_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

Hurdleポアソン回帰モデル

- Hurdleポアソン回帰モデルは、ロジスティック回帰モデルとポアソン回帰モデルを組み合わせたモデルとも言える
- ロジットリンクに対する線形予測子と対数リンクに対する線形予測子は、同じ説明変数のセットから組み合わせを変えて選ぶことができるが、便宜上説明変数をz_{ij}とx_{ij}に分けて表記する（同じ組み合わせでも構わない）

- π_iのロジットリンクと線形予測子

$$\mathrm{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = z_{i0}\delta_0 + z_{i1}\delta_1 + \cdots + z_{ik}\delta_k$$

- λ_iの対数リンクと線形予測子

$$\log(\lambda_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

Hurdleポアソン回帰モデル

- Hurdleポアソン回帰モデルの対数尤度は次式のようになる

$$\ln L(\pi_i, \lambda_i, y_i) = \begin{cases} \ln\pi_i, & y = 0 \\ \ln\left\{(1-\pi_i)\dfrac{{\lambda_i}^{y_i}}{(\exp(-\lambda_i)-1)y_i!}\right\}, & y = 1, 2, 3, \ldots \end{cases}$$

Hurdleポアソン回帰モデル

- $y=0$となる標本iのサブセットをΩ_0、 $y=1,2,3,\ldots$となる標本iのサブセットをΩ_1とすると、対数尤度関数は以下のように変形できる

$$\ln L(\pi,\lambda,y)=\ln\left\{\prod_{i\in\Omega_0}\pi_i\prod_{i\in\Omega_1}(1-\pi_i)\prod_{i\in\Omega_1}\frac{{\lambda_i}^{y_i}}{(\exp(-\lambda_i)-1)y_i!}\right\}$$
$$=\left\{\sum_{i\in\Omega_0}\ln\pi_i+\sum_{i\in\Omega_1}\ln(1-\pi_i)\right\}$$
$$+\sum_{i\in\Omega_1}\{y_i\ln\lambda_i-\ln(\exp(-\lambda_i)-1)-\ln(y_i!)\}$$

- 上の最終式における２つの{　}のうち、前者はδ_jの、後者はβ_jに関する未知パラメータに関する対数尤度関数となっている

ゼロ過剰ポアソン回帰モデル

- ゼロ過剰ポアソン回帰モデル（ゼロ強調ポアソン回帰モデル）は、次式のように表せる

$$P(Y_i=y_i)=\begin{cases}\pi_i+(1-\pi_i)\exp(-\lambda_i), y=0\\ (1-\pi_i)\dfrac{{\lambda_i}^{y_i}\cdot\exp(-\lambda_i)}{y_i!}, y=1,2,3,\ldots\end{cases}$$

ゼロ過剰ポアソン回帰モデル

- π_iはロジスティックリンクで、説明変数行列Zの要素z_{ij}と未知パラメータδ_jを用いて線形予測子は次式のように表せる

$$\pi_i = \frac{\rho_i}{1+\rho_i}$$

ここで、

$$\log(\rho_i) = z_{i0}\delta_0 + z_{i1}\delta_1 + \cdots + z_{ik}\delta_k$$

- λ_iは対数リンクで線形予測子は次式のようになる

$$\log(\lambda_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

ゼロ過剰ポアソン回帰モデル

- ゼロ過剰ポアソン回帰モデルの対数尤度は次式のようになる

$$\ln L(\pi_i, \lambda_i, y_i) = \begin{cases} \ln\{\pi_i + (1-\pi_i)\exp(-\lambda_i)\}, y = 0 \\ \ln\left\{(1-\pi_i)\dfrac{{\lambda_i}^{y_i}\cdot\exp(-\lambda_i)}{y_i!}\right\}, y = 1, 2, 3, \ldots \end{cases}$$

ゼロ過剰ポアソン回帰モデル

- $y=0$となる標本iのサブセットをΩ_0、 $y=1,2,3,\ldots$となる標本iのサブセットをΩ_1とすると、対数尤度関数は以下のように変形できる

$$\begin{aligned}
&\ln L(\pi_i,\lambda_i,y_i)\\
&=\ln\left\{\prod_{i\in\Omega_0}\{\pi_i+(1-\pi_i)\exp(-\lambda_i)\}\prod_{i\in\Omega_1}(1-\pi_i)\prod_{i\in\Omega_1}\frac{{\lambda_i}^{y_i}\cdot\exp(-\lambda_i)}{y_i!}\right\}\\
&=\sum_{i\in\Omega_0}\ln[\rho_i+\exp(-\lambda_i)]+\sum_{i\in\Omega_1}\{y_i\ln\lambda_i-\lambda_i-\ln(y_i!)\}\\
&+\sum_{i=1}^{n}\ln\{1+\rho_i\}
\end{aligned}$$

負の二項分布

- 負の二項分布のprobability mass functionは次式の通り

$$f_{NB}(\,y|\lambda,\phi)=\frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)}\left(\frac{\phi}{\phi+\lambda}\right)^{\phi}\left(\frac{\lambda}{\phi+\lambda}\right)^{y}$$

- このとき

平均： λ

分散： $\frac{\lambda^2}{\phi}+\lambda$

- $(1+\phi)/\phi$を分散指標、 ϕをscale parameterという
- $\phi\to+\infty$としたとき、負の二項分布はポアソン分布となる

負の二項分布モデル

- 被説明変数が負の二項分布に従うとする負の二項分布モデルは、例えば次式のように表される。ここでy_iは被説明変数、X_iは説明変数、$\boldsymbol{\beta}$は未知パラメータである

$$y_i = NB(\lambda_i)$$
$$f_{NB}(y_i|\lambda_i,\phi) = \frac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\left(\frac{\phi}{\phi+\lambda_i}\right)^{\phi}\left(\frac{\lambda_i}{\phi+\lambda_i}\right)^{y_i}$$
$$g\big(E(y_i)\big) = \lambda_i = \exp(\eta_i)$$

- このときlogリンク(対数リンク)$\log(\lambda_i)$と線形予測子$\eta_i = X_i\boldsymbol{\beta}$との関係は、次式のように対数リンク関数で表せる

$$\log(\lambda_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = X_i^T\boldsymbol{\beta}$$

Hurdle負の二項分布モデル

- Hurdle負の二項分布モデルは、次式のように表せる

$$P(Y_i = y_i) = \begin{cases} \pi_i, & y = 0 \\ (1-\pi_i)\dfrac{\Gamma(y_i+\phi)}{\Gamma(y_i+1)\Gamma(\phi)}\left(\dfrac{\phi}{\phi+\lambda_i}\right)^{\phi}\left(\dfrac{\lambda_i}{\phi+\lambda_i}\right)^{y_i}, & y = 1,2,3,\ldots \end{cases}$$

Hurdle負の二項分布モデル

- π_iはロジットリンクで、説明変数行列Zの要素z_{ij}と未知パラメータδ_jを用いて線形予測子は次式のように表せる

$$\mathrm{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = z_{i0}\delta_0 + z_{i1}\delta_1 + \cdots + z_{ik}\delta_k$$

- λ_iの対数リンクと線形予測子は次式のようになる

$$\log(\lambda_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

ゼロ過剰負の二項分布モデル

- ゼロ過剰負の二項分布モデルは、次式のように表せる

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i) f_{NB}(y_i = 0 | \lambda_i, \phi), y = 0 \\ (1-\pi_i) f_{NB}(y_i > 0 | \lambda_i, \phi), y = 1, 2, 3, \ldots \end{cases}$$

$$f_{NB}(y_i | \lambda_i, \phi) = \frac{\Gamma(y_i + \phi)}{\Gamma(y_i + 1)\Gamma(\phi)} \left(\frac{\phi}{\phi + \lambda_i}\right)^{\phi} \left(\frac{\lambda_i}{\phi + \lambda_i}\right)^{y_i}$$

ゼロ過剰負の二項分布モデル

- π_iはロジスティックリンクで、説明変数行列Zの要素z_{ij}と未知パラメータδ_jを用いて線形予測子は次式のように表せる

$$\pi_i = \frac{\rho_i}{1+\rho_i}$$

ここで、

$$\log(\rho_i) = z_{i0}\delta_0 + z_{i1}\delta_1 + \cdots + z_{ik}\delta_k$$

- λ_iは対数リンクで線形予測子は次式のようになる

$$\log(\lambda_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

ゼロ過剰負の二項分布モデル

- ゼロ過剰負の二項分布モデルの対数尤度は次式のようになる

$$\ln L(\pi_i, \lambda_i, y_i) = \begin{cases} \ln\{\pi_i + (1-\pi_i) f_{NB}(y_i|\lambda_i, \phi)\}, y = 0 \\ \ln\{(1-\pi_i) f_{NB}(y_i|\lambda_i, \phi)\}, y = 1, 2, 3, \ldots \end{cases}$$

ゼロ過剰負の二項分布モデル

- $y=0$となる標本iのサブセットをΩ_0、 $y=1,2,3,\ldots$となる標本iのサブセットをΩ_1とすると、対数尤度関数は以下のように変形できる

$$
\begin{aligned}
&\ln L(\pi_i, \lambda_i, y_i) \\
&= \ln\left\{\prod_{i\in\Omega_0}\{\pi_i + (1-\pi_i) f_{NB}(y_i|\lambda_i, \phi)\}\prod_{i\in\Omega_1}(1-\pi_i) f_{NB}(y_i|\lambda_i, \phi)\right\} \\
&= \sum_{i\in\Omega_0}\ln\left[\rho_i + \left(\frac{\phi}{\phi+\lambda_i}\right)^{\phi}\right] \\
&+ \sum_{i\in\Omega_1}\{\ln[\Gamma(y_i+\phi)] - \ln[\Gamma(y_i+1)] - \ln[\Gamma(\phi)] + \phi\ln\phi + y_i\ln\lambda_i \\
&- (\phi + y_i)\ln(\phi+\lambda_i)\} \\
&+ \sum_{i=1}^{n}\ln\{1+\rho_i\}
\end{aligned}
$$