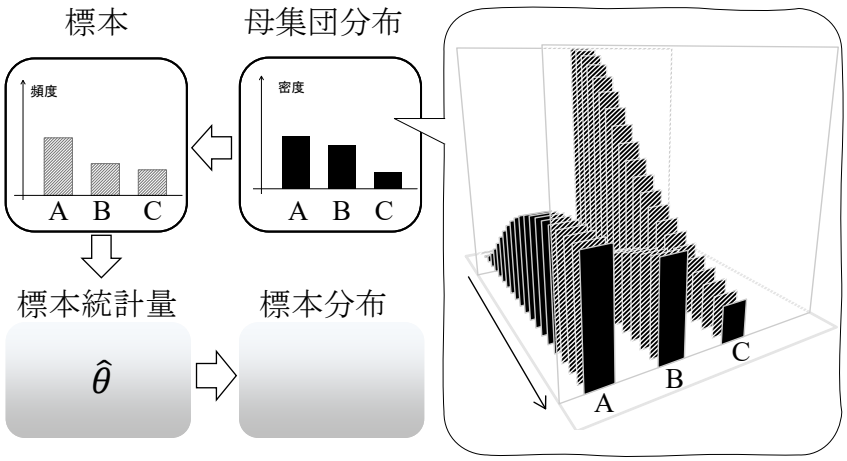
第3講　一般化線形モデル２：ポワソン回帰

学びのポイント

- 粗頻度を従属変数とし、その値を独立変数たちから予測するモデルの一つに、**ポワソン回帰**が存在することが分かる。
- ポワソン回帰の構造を**数式**で表現することができる。
- ポワソン回帰のパラメータは負の値を取らず、この制約を満たすため、ポワソン回帰のリンク関数には**対数リンク**を用いることが分かる。
- 重回帰モデル、ロジスティック回帰モデル、およびポワソン回帰モデルは、独立変数の線形結合と、**母集団確率分布の平均**を、リンク関数を介してつなげているという点で共通した構造を持ち、このような構造を持つ統計モデルを**一般化線形モデル**と呼ぶことが分かる。
- 「単位○○あたりの頻度」の期待値をモデル化したいときには、**オフセット項つきのポワソン回帰**を用いることが分かる。

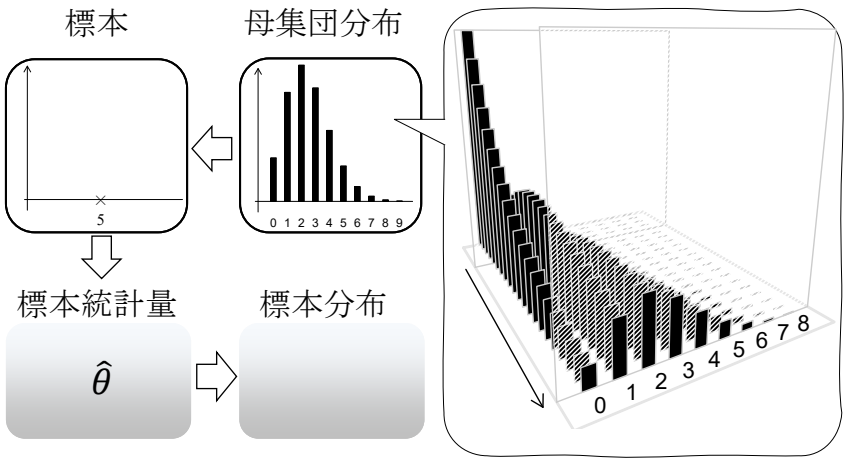
見取り図

【第２講】ロジスティック回帰モデル



【第３講】ポワソン回帰モデル

☞ノート０：基本的な考え方
☞ノート１：統計モデル
☞ノート２：統計モデル（オフセット項）



■ 目標

第 2 講であつかったロジスティック回帰モデルは、リンク関数を用いて、独立変数の線形結合をベルヌーイ分布／二項分布の期待値パラメータと結びつける統計モデルでした。このように、リンク関数を母集団確率分布の期待値に結びつける構造を持つ統計モデルを一般化線形モデルと呼びます。

この第 3 講では、一般化線形モデルの別のモデルとしてポワソン回帰を学びます。これは、ポワソン分布の期待値パラメータλと独立変数たちを、対数リンク関数で結び付けたものです。

同じ一般化線形モデルということで、推定方法も最尤推定法を用いたりと、基本的な統計推論のパターンは第 2 講のものと全く同じです。そこで、この第 3 講では、重複する部分は割愛し、ポワソンモデルの特徴である統計モデルの作り方に焦点を当て、どのような考え方で、このモデルを立て、使用するのかに注目して議論を進めます。

🕮 ノート０　ポワソン回帰の目的と基本的な考え方

○目的

回帰分析と同じやりかたで、

従属変数がカウントデータの場合も扱いたい！

イメージ

	点推定値	標準誤差	p 値
切片	1.20	0.11	0.00 ***
回帰係数１	- 2.11	0.21	0.00 ***
回帰係数２	0.80	0.33	0.02 *

音韻論における利用例

① モチベーション

それぞれの言語にいくつの音素が存在するか予測したい！

② 例

それぞれの言語が持つ音素の数を (a) その言語が属する語族、(b) 話者の人口等から予測する。

第二言語教育における利用例

① モチベーション

英語を学んでいる学習者が間違いを起こす回数を、その学習者の特徴から予測、説明したい！

② 例

自由作文を実施し、その中で英語のスペルを間違えた数を (a) 学習者のレベル、(b) 作文のテーマなどから予測する。

社会言語学における利用例

① モチベーション

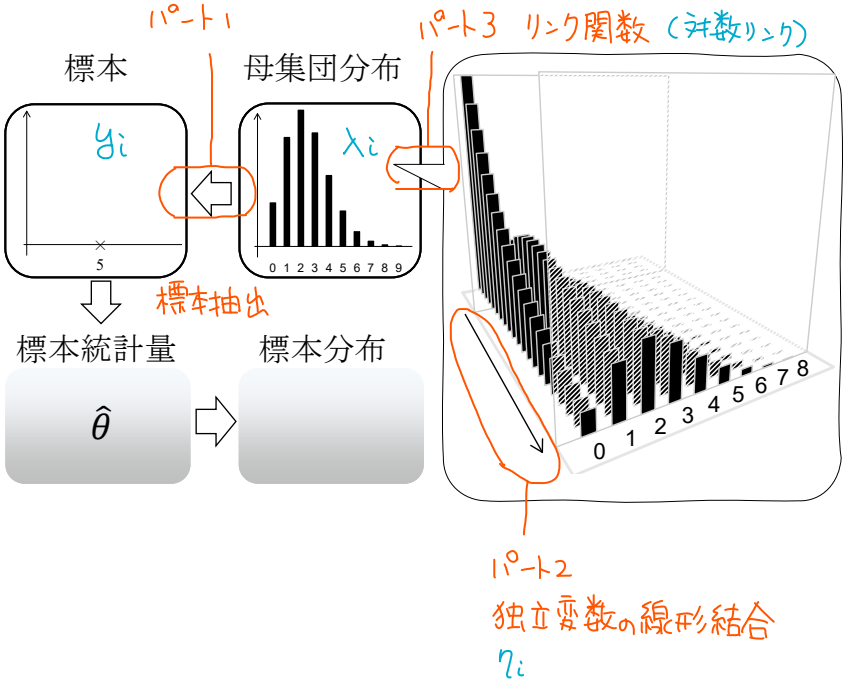
フィラーが生産される回数を社会要因等から説明したい！

② 例

「えー／あー」などのフィラーの生産回数を (a) 発話時のフォーマルさ、(b) 話者の性別、(c) 話者の社会階層等から予測する。

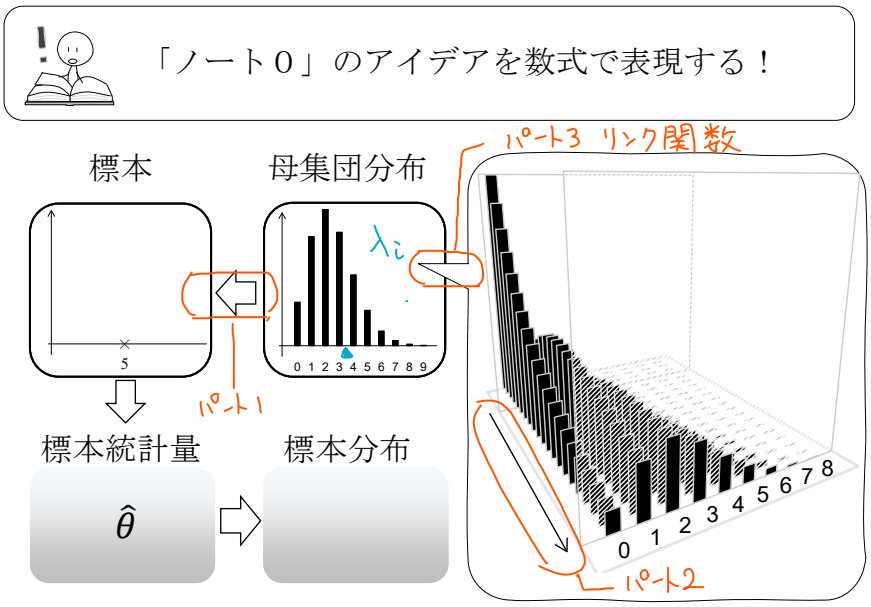
○ ポイント

独立変数をもとに母集団の確率を　構造化　する！



🕮 ノート１　母集団分布：統計モデル

○目的



（１）パート１：母集団分布からの標本抽出

$$y_i \sim Po(\lambda_i)$$

$0 < \lambda_i$

（２）パート２：独立変数たちの線形結合を作る

$-\infty \leq \eta_i \leq \infty$

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni}$$

（３）パート３：リンク関数 Link Function

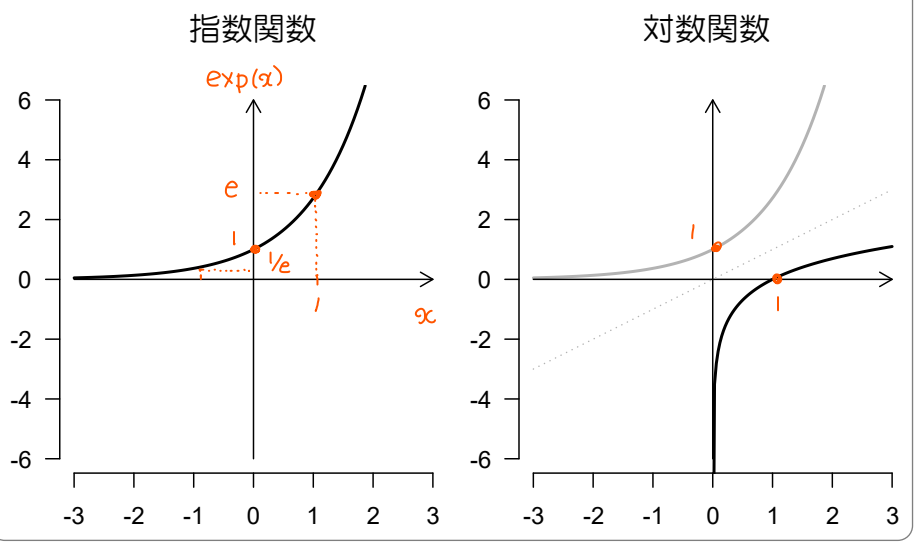
$(-\infty, \infty)$の値を取るη_iから、$(0, \infty)$の値を取るλ_iへ、(a) 一対一で、(b)滑らかに、変換する関数のこと。

$$\lambda_i = F(\eta_i) \qquad \lambda_i = \exp[\eta_i]$$

$$G(\lambda_i) = \eta_i \qquad \log(\lambda_i) = \eta_i$$

ポワソン分布では、F に指数関数。G に対数関数を用いる。

指数関数と対数関数



まとめ：一般化線形モデル

① ポワソン回帰

$$y_i \sim Po(\lambda_i)$$

$$\lambda_i = exp(\eta_i)$$

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni}$$

② ロジスティック回帰

$$y_i \sim Bern(\pi_i)$$

$$\pi_i = \frac{exp(\eta_i)}{1 + exp(\eta_i)}$$

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni}$$

③ 重回帰

$$y_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \eta_i$$

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni}$$

※　前期に習った重回帰分析はリンク関数が恒等関数（Identity function）である場合だと考えられる。

📖 ノート２　母集団分布：オフセット項

○目的

「頻度」ではなく、「単位○○あたりの頻度」を
独立変数たちから予測したい！

ID（コーパス）	y_i	x_{1i}（性別）	x_{2i}（講演）	a_i（収録時間）
1	2	0	1	30 分
2	3	0	1	45 分
3	0	0	0	30 分
4	11	0	0	300 分
5	9	1	0	400 分
6	7	0	1	200 分
⋮	⋮	⋮	⋮	⋮

（１）ふつうのポワソン回帰：粗頻度をモデル化

$$\begin{cases} y_i \sim Po(\lambda_i) \\ \lambda_i = \exp(\eta_i) \\ \eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni} \end{cases}$$

（２）オフセット項つき：単位○○あたりの粗頻度をモデル化

「単位○○あたりの平均頻度 $\frac{\lambda_i}{a_i}$」をモデル化する。

$$\begin{cases} y_i \sim Po(\lambda_i) \\ \lambda_i = \exp(\eta_i) \\ \eta_i = \underbrace{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni}}_{\text{よく知っている構造}} + \underbrace{\log(a_i)}_{\text{おまけ（offset）}} \end{cases}$$

Ⓟ 指数の計算

(1) 簡単なケース

$2^3 \times 2^5 = 2^{3+5}$

(2) 少し難しくしたケース

$2^3 \times 3 = 2^{3+\log_2 3}$

$\left(2^3 \times 2^{\log_2 3} = 2^{3+\log_2 3}\right)$

(3) 少しさらに難更

$e^3 \times e^5 = e^8$

$\exp(3) \times \exp(5) = \exp(8)$

$\exp(3) \times 3 = \exp(3+\log 3)$

$\exp(\log 3)$

Ⓟ 記法 : exp

(1) ネイピア数

$e = 2.718\cdots$

(2) 指数関数

$\exp(x)$

= ネイピア数 e を x 乗した値

オフセット項が出てくるわけ

単位時間あたりの期待頻度

$$\begin{cases} y_i \sim Po(\lambda_i) \\ \dfrac{\lambda_i}{a_i} = \exp(\eta_i') \\ \eta_i' = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni} \end{cases}$$

$$\Leftrightarrow \begin{cases} y_i \sim Po(\lambda_i) \\ \lambda_i = \exp(\eta_i') \times a_i \\ \quad\ = \exp(\eta_i' + \log(a_i)) \\ \eta_i' = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni} \end{cases}$$

ここで、$\eta_i = \eta_i' + \log(a_i)$ と置き換えると、

$$\Leftrightarrow \begin{cases} y_i \sim Po(\lambda_i) \\ \lambda_i = \exp(\eta_i) \\ \eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni} + \log(a_i) \end{cases}$$

オフセット項の解釈

$$\begin{aligned} \eta_i &= \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni} + \log(a_i) \\ &= \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni} + x_{(n+1)i} \end{aligned}$$

$x_{(n+1)i} = \log(a_i)$ とおけば、オフセット項は「偏回帰係数が常に 1 に固定化された独立変数だ」と解釈できる。

オフセット項の使いどころ

左ページでは、「単位時間あたり」の例を挙げたが、「単位面積あたりの何かの出現頻度」のように様々な応用が可能。

R における実装

```
fit <- glm(y ~ x1 + x2, offset = log(a), data = data, family = poisson)
```