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

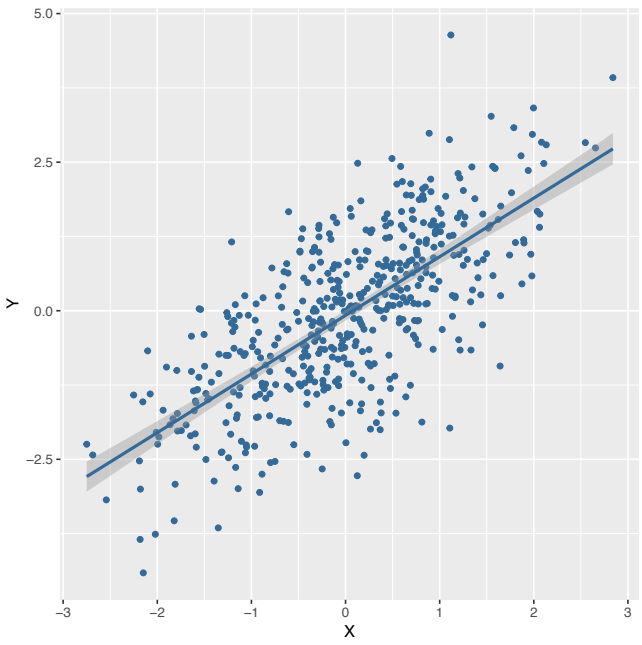
代表的な一般化線形回帰モデル

- 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

線形回帰モデル

- 説明変数と被説明変数がともに正規分布
- 誤差項も正規分布

- 説明変数と被説明変数との関係が線形式で表される



線形回帰モデル（重回帰分析）

- 従属変数yとk個の独立変数$x_1, x_2, \ldots, x_k$に対する標本数がn個の重回帰モデルは以下のように記述できる$(i = 1, \ldots, n)$

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\
&\vdots \\
y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_k x_{nk} + \varepsilon_n
\end{aligned}
$$

線形回帰モデル（重回帰分析）

- 次のようなベクトルと行列を用いて、

$$
\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \cdots & x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}
$$

- 次式のように簡略化できる

$$
\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}
$$

あるいは

$$
y_i = X_i^T \boldsymbol{\beta} + \varepsilon_i
$$

線形回帰分析を行う上での仮定（前提）

- 線形回帰分析では、独立変数と従属変数がともに正規分布に従うことを前提としている

- 独立変数行列Xが平均μ、分散Σの正規分布に従う$X \sim N(\mu, \Sigma)$とき、$X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \sim N(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\beta}\Sigma\boldsymbol{\beta}^T)$となる

- さらに誤差項$\boldsymbol{\varepsilon}$が平均0、分散σ^2の正規分布に従う$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$と仮定している。
- このことから従属変数$\boldsymbol{y}$は平均$X\boldsymbol{\beta}$、分散$\sigma^2 I$の正規分布に従う

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \sim N(X\boldsymbol{\beta}, \sigma^2 I)$$

一般化線形モデル

- 線形回帰モデルでは、説明変数と被説明変数がともに正規分布に従い、誤差項が互いに独立で同一の正規分布に従うと仮定
- しかし、すべてのデータが正規分布に従うとは限らない
- 被説明変数が正規分布に従わない時、$E(\boldsymbol{y}) = X\boldsymbol{\beta}$と仮定するとモデルの正確さが失われる

- データが正規分布以外の確率分布に従い、説明変数と被説明変数との関係をリンク関数と線形予測子を用いて推定するモデルを一般化整形モデルという

一般化線形モデル

- 被説明変数yと、k個の説明変数$x_1, x_2, \ldots, x_k$に対する標本数がn個の一般化線形モデルは以下のように記述できる$(i = 1, \ldots, n)$

$$y_i \sim f(y_i|\theta)$$
$$g\big(E(y_i)\big) = g(\mu) = \eta_i$$
$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = X_i^T \boldsymbol{\beta}$$

- $f(y_i|\theta)$は被説明変数が従う確率分布の確率密度関数であり、θはその確率密度関数のパラメータ
- $g\big(E(y_i)\big) = g(\mu) = \eta_i$をリンク関数
- $\eta_i = X_i^T \boldsymbol{\beta}$を線形予測子という

一般線形モデルの最尤推定

- 一般線形モデルの尤度関数$L(\theta|y)$は以下のようになる

$$L(\theta|y_1, \ldots, y_2, \ldots, y_n) = \prod_{i=1}^{n} L(\theta|y_i)$$

- 尤度関数を解析的に解くことは難しいため、尤度関数に対数をとった対数尤度関数の最適解（最大値）を求めることにより、未知パラメータを計算する

一般化線形モデルの確率分布とリンク関数

モデル	被説明変数	確率分布	リンク関数
線形回帰モデル	実数	正規分布	恒等リンク
ロジスティック回帰モデル	0/1の二値	二項分布	logitリンク
プロビット回帰モデル	0/1の二値	二項分布	probitリンク
補対数対数モデル	0/1の二値	二項分布	cloglogリンク
ポアソン回帰モデル	非負の整数	ポアソン分布	logリンク
負の二項分布モデル	非負の整数	負の二項分布	logitリンク
ベータ回帰モデル	[0,1]の実数	ベータ分布	logitリンク
ガンマ回帰モデル	非負の実数	ガンマ分布	逆logリンク
指数-ガウス回帰モデル	裾の長い実数	指数-ガウス分布	logリンクor 恒等リンク

0/1の二値を被説明変数とする回帰モデル

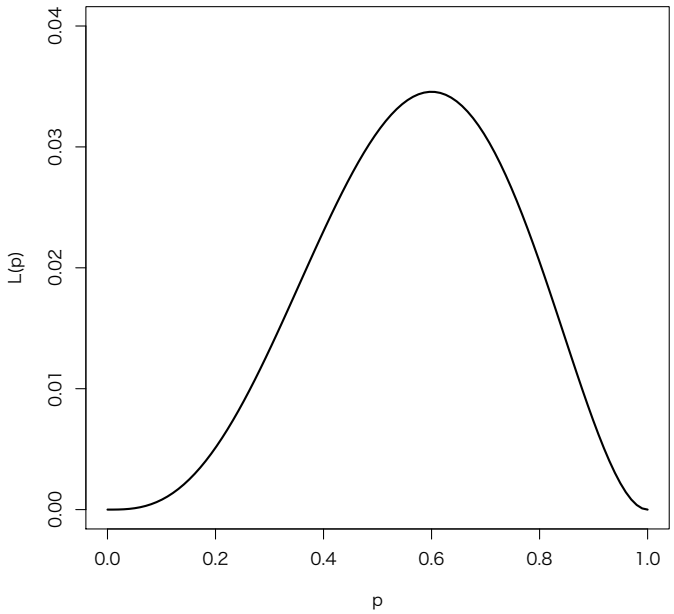
- 適用される主なモデル
  - ロジスティック回帰モデル
  - プロビット回帰モデル
  - 補体数対数モデル
- 適用例
  - 病気の発生有無、健康であるかどうか、リスク発生の有無
  - 店舗選択の有無、商品購入の有無（意向）
  - 試験の合否、ストレスの有無、

ベルヌーイ試行と二項分布

- 0か1かしかない試行において、n回の試行でr回成功し、その確率πがわかっているとき、実験が成功する期待値はベルヌーイ試行に従う
- ベルヌーイ試行の確率分布を二項分布といい、その分布は次式の確率密度関数に従う

$$Binom(n,p) = {}_nC_r \cdot \pi^r \cdot (1-\pi)^{n-r} \approx \pi^r \cdot (1-\pi)^{n-r}$$

二項分布の例



二項分布の尤度

- $y_i = 1$のときの確率をπ_i、 $y_i = 0$のときの確率を$1 - \pi_i$とする
- 二項分布の尤度L_iは次式のようになる

$$L_i = {\pi_i}^{y_i} \cdot (1 - \pi_i)^{1 - y_i}$$

- 従って尤度関数Lは以下のように表せる

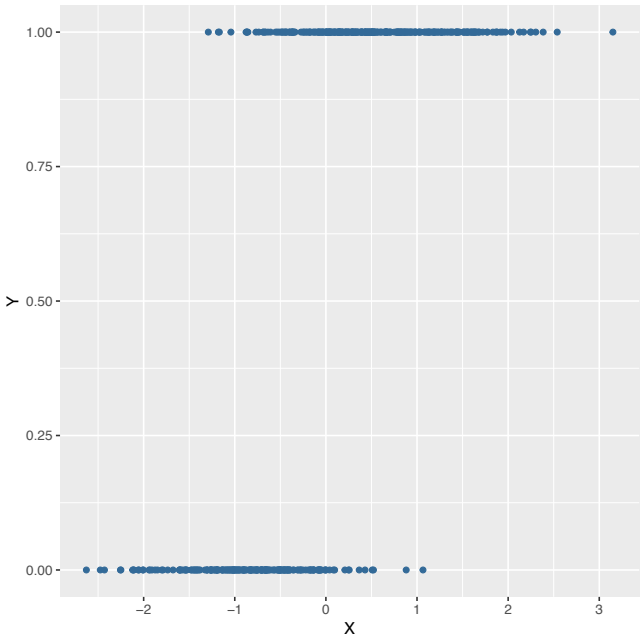
$$L = \prod_{i=1}^{n} {\pi_i}^{y_i} \cdot (1 - \pi_i)^{1 - y_i}$$

- また対数尤度関数は以下のようになる

$$\ln L = \sum_{i=1}^{n} \left(y_i \cdot \ln \pi_i + (1 - y_i) \cdot \ln(1 - \pi_i) \right)$$

ロジスティック回帰モデル

- 被説明変数は0または1の値を取る
- 被説明変数の例
  - 商品購入の有無
  - 投票の有無
  - 賛成/反対
  - 疾患の有無
  - 事象発生の有無



（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

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

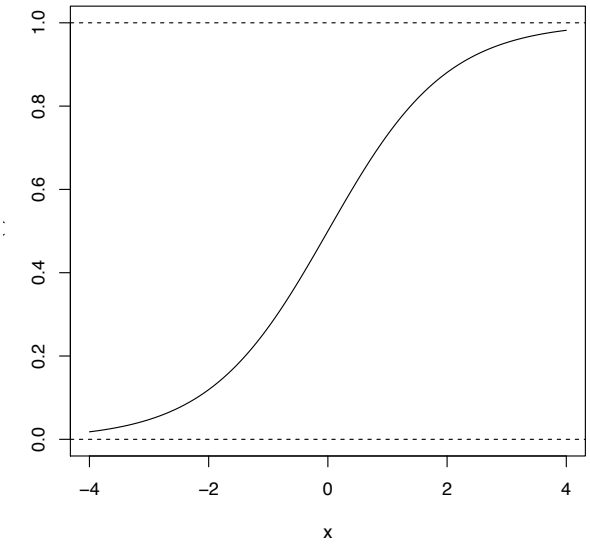
（二項）ロジスティック回帰モデル

- $\frac{\pi_i}{1-\pi_i}$はオッズ(odds)比
- 「オッズが1より大きい」は「$\pi_i > 0.5$」と同じ
- $\ln\left(\frac{\pi_i}{1-\pi_i}\right)$は対数オッズまたはロジットという
- $\mathrm{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$をlogitリンク（ロジットリンク）という

- イベントやリスクの発生確率を説明するモデルなどに用いられる
- y_iを確率値に変換して推定することも可能
- 二項ロジットモデル(binomial logit model)とも呼ばれる

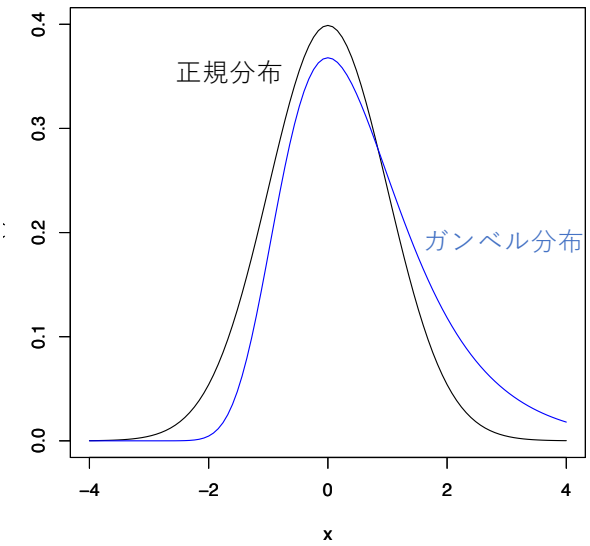
ロジスティック分布の累積密度関数

- ロジスティック分布の累積密度関数は、右図のように0-1の間の値を取る
- ロジスティック回帰モデルの被説明変数としてロジスティック分布の累積密度関数の確率値を用いることもできる



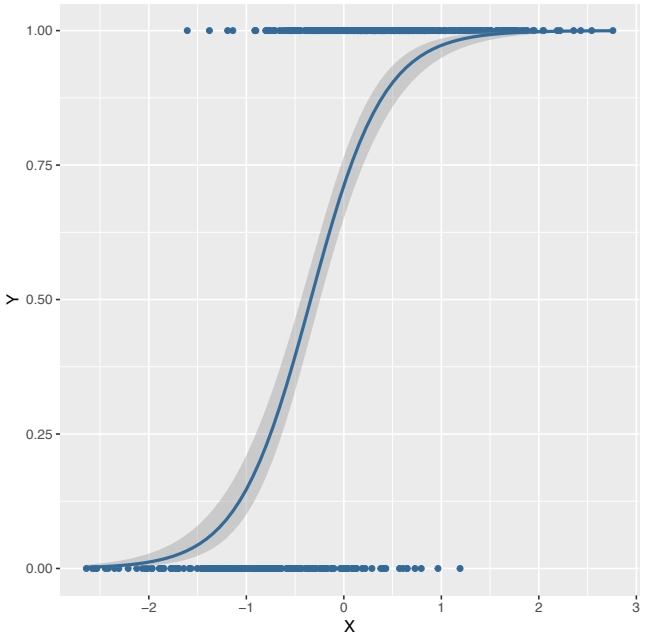
（二項）ロジスティック回帰モデル

- （二項）ロジスティック回帰モデル（二項ロジットモデル）の誤差項ε_iは、ガンベル分布に従う
- 正規分布とガンベル分布の違いは右図の通り



ロジスティック回帰モデルの推定結果の例

- 500個のランダムな二項分布に従う被説明変数と正規分布に従う説明変数を用いてロジスティック回帰モデルを推定
- 得られたモデルの曲線と95%信頼区間は右図のとおり



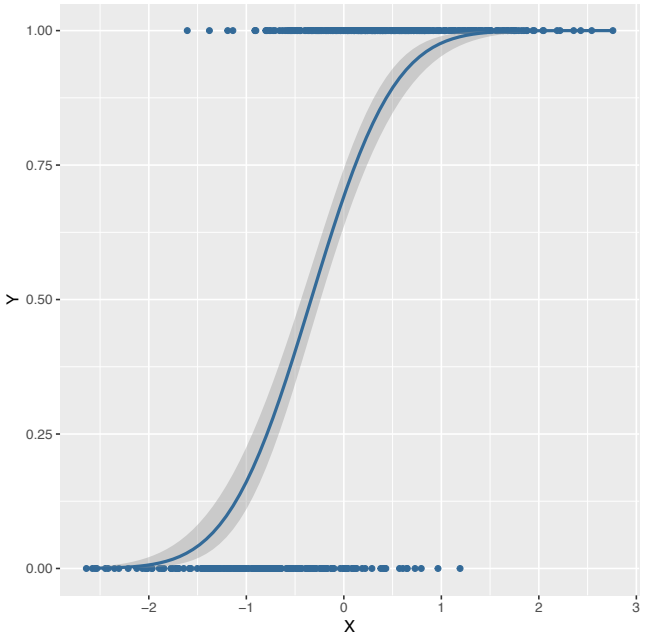
（二項）プロビットモデル

- 二項ロジスティック回帰モデル（二項ロジットモデル）では、π_iにロジスティック分布の確率密度関数$\pi_i = \frac{1}{1+exp\left(-\left(X_i^T\boldsymbol{\beta}\right)\right)}$を考えた
- π_iに標準正規分布の確率密度関数を与えたものを、二項プロビットモデルという

$$\pi_i = \int_{-\infty}^{X_i^T\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y^2}{2}\right) dy$$

プロビットモデルの推定結果の例

- 500個のランダムな二項分布に従う被説明変数と正規分布に従う説明変数を用いてプロビット回帰モデルを推定
- 得られたモデルの曲線と95%信頼区間は右図のとおり



補対数対数モデル

- 0～1までの確率π_iについて、補対数対数関数（cloglog関数）は以下のように定義される

$$\mathrm{cloglog}(\pi_i) = \log\big(-\log(1-\pi_i)\big)$$

- リンク関数に補対数対数リンク関数を与えた一般化線形回帰モデルを、補対数対数モデルという

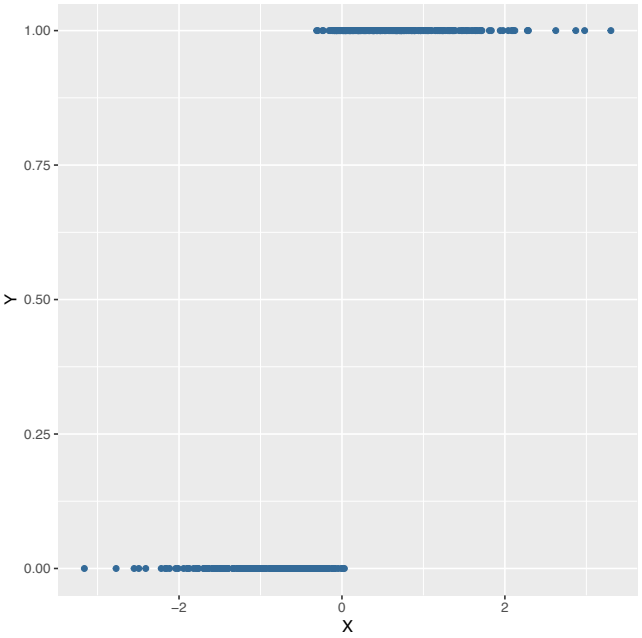
$$\log\big(-\log(1-\pi_i)\big) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = X_i^T \boldsymbol{\beta}$$

- ここで補対数対数モデルのリンクは次式のようになる

$$g\big(E(y_i)\big)\pi_i = 1 - \exp\big(\exp(\eta_i)\big)$$

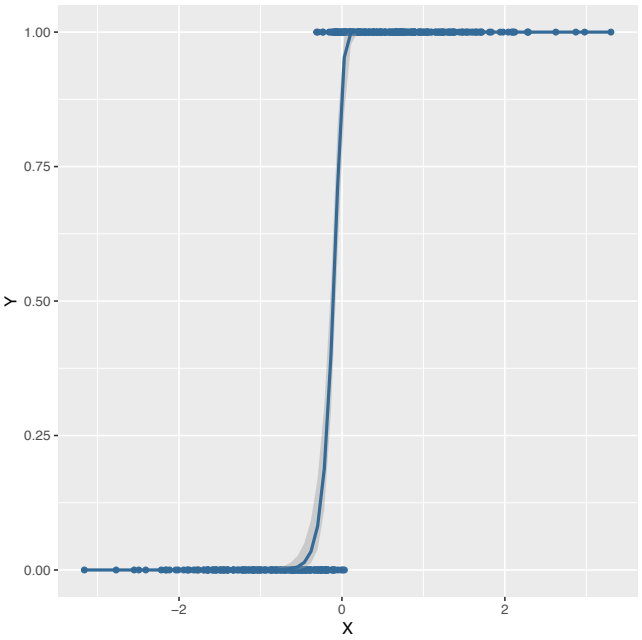
補対数対数モデル

- ロジスティック回帰モデルやプロビット回帰モデルは左右対称なリンク関数を用いる
- 補対数対数モデルはリンク関数が非対称性をもたない観測度数や共変量の場合に用いられる



補対数対数モデルの推定結果の例

- 500個のランダムな補対数対数分布に従う被説明変数と正規分布に従う説明変数を用いてロジスティック回帰モデルを推定
- 得られたモデルの曲線と95%信頼区間は右図のとおり



ロジスティック回帰モデルと補対数対数モデルの違い

- ロジスティック回帰モデル
  - 推定された$\exp(\beta_k)$は、説明変数x_kの 1 単位の変化に対するオッズ比の変化を意味する
  - リンク関数は対称

- 補対数対数モデル
  - 推定された$\exp(\beta_k)$は、説明変数x_kの 1 単位の変化に対するハザード確率の変化を意味する
  - リンク関数は非対称