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

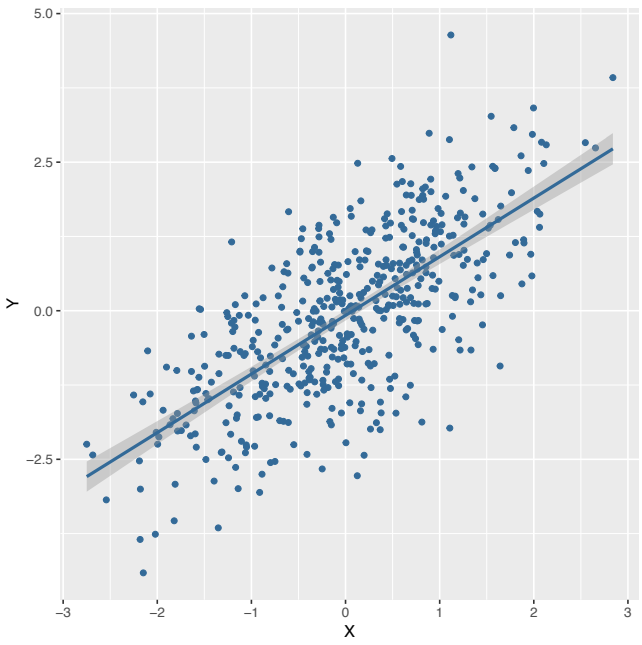
代表的な一般化線形回帰モデル

- 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$(i = 1, \ldots, n)$

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

- $f(y_i|\theta)$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
- $g\big(E(y_i)\big) = \mu$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

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s回成功し、その確率pがわかっているとき($s = n \times p$)、実験が成功する期待値は以下の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

$$Binom(n,p) = {}_nC_s \cdot p^s \cdot (1-p)^{n-s} \propto p^s \cdot (1-p)^{n-s}$$

二項分布

- $s = \alpha - 1, n - s = \beta - 1$とすると、二項分布$Binom(n, p)$は次式のように変形できる

$$\begin{aligned} Binom(n,p) &\propto p^s \cdot (1-p)^{n-s} \\ &= p^{\alpha-1} \cdot (1-p)^{\beta-1} \\ &= \frac{(\alpha-1)!\,(\beta-1)!}{(\alpha+\beta-1)!} \end{aligned}$$

ベータ分布

- 二項分布の式変形の結果から、ベータ分布の確率密度関数$f(\alpha, \beta, p)$を関係づけることができる

$$\begin{aligned} f(p;\alpha,\beta) = Beta(\alpha,\beta) &= k \cdot p^{\alpha-1} \cdot (1-p)^{\beta-1} \\ &= k \cdot \frac{(\alpha-1)!\,(\beta-1)!}{(\alpha+\beta-1)!} \end{aligned}$$

$$0 < p < 1, 0 < \alpha, 0 < \beta$$

$$k = \frac{1}{\int_0^1 p^{\alpha-1} \cdot (1-p)^{\beta-1} dp}$$

ベータ分布

- ベータ分布$Beta(\alpha,\beta)$の平均μと分散σ^2は以下のようになる

$$\mu = \frac{\alpha}{\alpha+\beta}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

ベータ関数

- 一般に、次式で表される関数をベータ関数$B(\alpha,\beta)$という

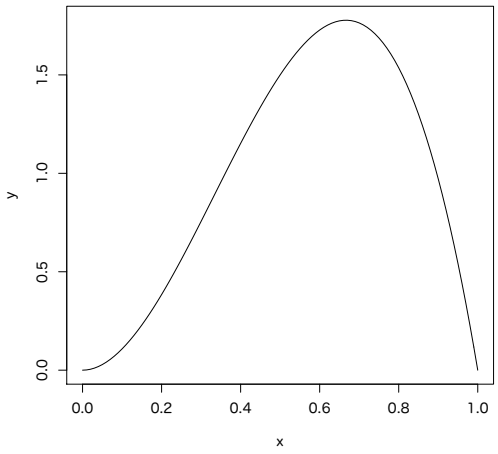
$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}\cdot(1-x)^{\beta-1}dx$$

- ベータ関数は以下の性質を持つ

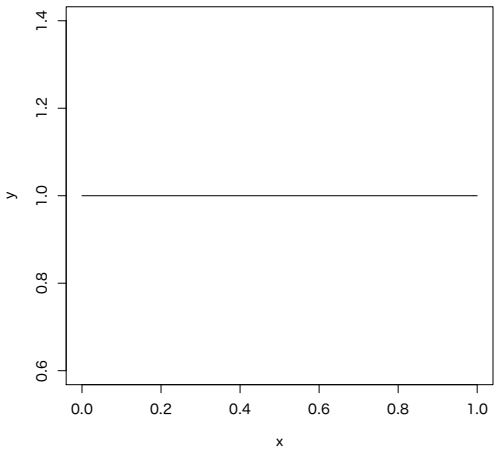
$$B(\alpha,\beta) = B(\beta,\alpha)$$
$$\alpha B(\alpha,\beta+1) = \beta B(\alpha+1,\beta)$$
$$B(\alpha,\beta) = B(\alpha,\beta+1) + B(\alpha+1,\beta)$$

ベータ分布

$B(3,2)$

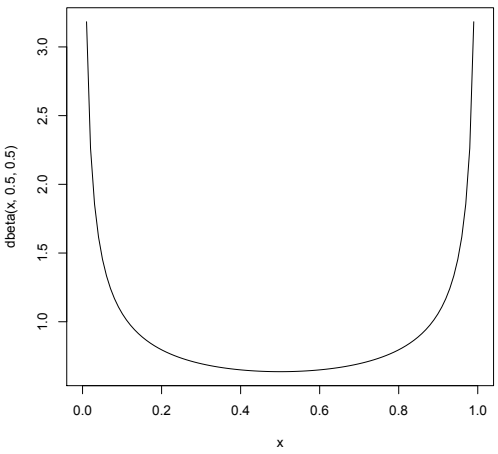


$B(1,1)$=一様分布

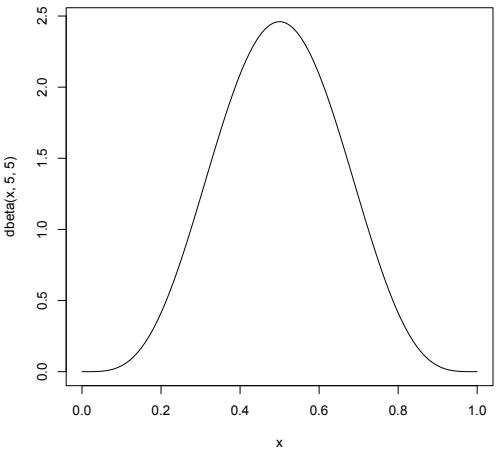


ベータ分布

$B(0.5,0.5)$



$B(5,5)$



ベータ分布は万能な確率分布

- ベータ分布は以下の分布に変化できる
  - 一様分布、線形分布
  - 単調増加・単調減少分布
  - 単峰分布
  - 左右対称分布

ベータ分布とガンマ分布

- ベータ関数の確率密度関数は以下のように式変形できる

$$Binom(n, p; \alpha, \beta) \approx p^{\alpha} \cdot (1-p)^{\beta-1}$$
$$= \frac{(\alpha-1)!\,(\beta-1)!}{(\alpha+\beta-1)!}$$
$$= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- ここで$\Gamma(\alpha)$はガンマ関数という

ベータ関数とガンマ関数

- ベータ関数は以下のように式変形できる

$$B(\alpha,\beta) = \int_0^1 p^{\alpha-1}\cdot(1-p)^{\beta-1}dp$$
$$= \frac{(\alpha-1)!\,(\beta-1)!}{(\alpha+\beta-1)!}$$
$$= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- ここで$\Gamma(\alpha)$はガンマ関数という

ガンマ関数

- ガンマ関数$\Gamma(\alpha)$は次式で表される

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx$$
$$\alpha > 0$$

- ガンマ関数$\Gamma(\alpha)$は以下のような性質を持つ

$$\Gamma(\alpha) = (\alpha-1)!$$
$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$$
$$\Gamma(1/2) = \sqrt{\pi}$$

ガンマ分布

- ガンマ分布の確率密度関数$f(x)$は次式のようになる

$$f(x) = Ga(\alpha, x) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, x \geq 0$$

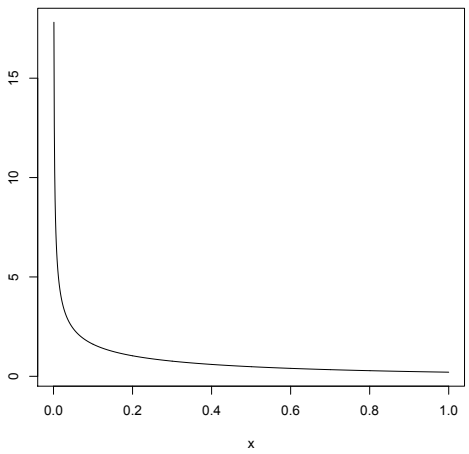
- ガンマ分布の平均$E(X)$と分散$V(X)$は以下の通り

$$E(X) = \frac{\alpha}{\lambda}$$

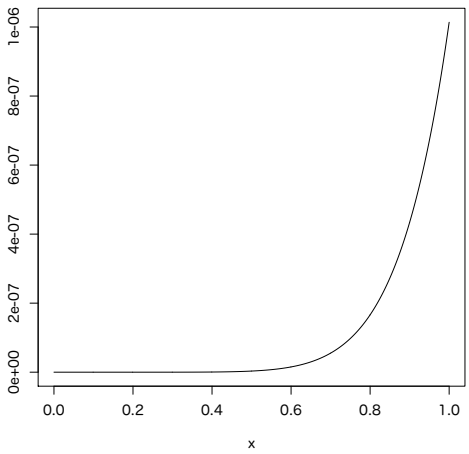
$$E(X) = \frac{\alpha}{\lambda^2}$$

ガンマ分布

$\Gamma(0.5)$



$\Gamma(10)$



指数分布

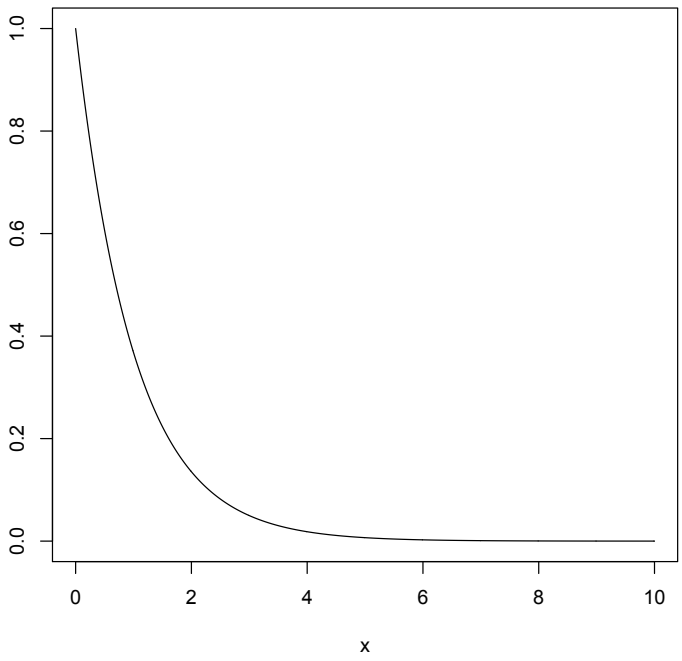
- ガンマ分布の確率密度関数について、$\alpha = 1$とすると指数関数となる

$$f(x) = \lambda e^{-\lambda x}, x \geq 1$$

$$E(x) = \frac{1}{\lambda}$$

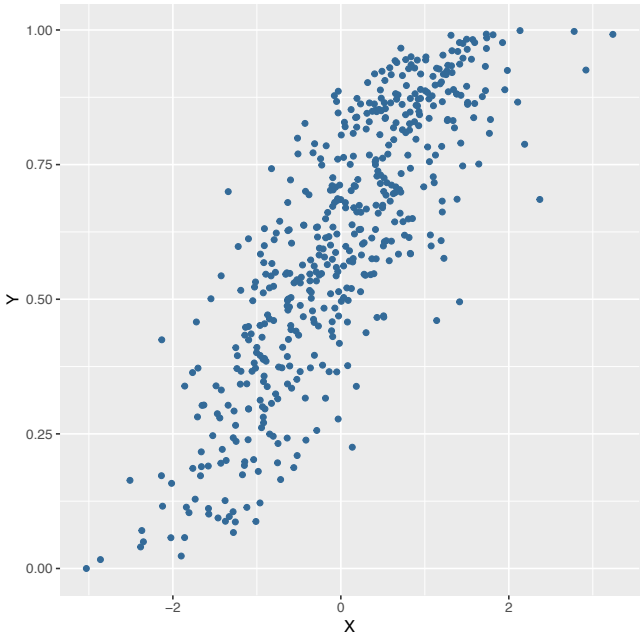
$$V(x) = \frac{1}{\lambda^2}$$

指数分布



ベータ回帰モデル

- 被説明変数yが0～1の間をとり、説明変数xとの間に右図のような関係が見られるような場合、ベータ回帰モデルが適用されることがある



ベータ回帰モデル

- ベータ回帰モデルでは、被説明変数yがベータ分布に従うと仮定する
- 被説明変数yが0～1の間をとる場合に用いられることがあるモデル

$$y \sim B(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot y^{\alpha-1} \cdot (1-y)^{\beta-1}, \qquad 0 \le y \le 1$$

ベータ回帰モデル

- 被説明変数yが従うベータ関数$y \sim B(\alpha, \beta)$は、被説明変数の平均$E(y) = \mu$と分散$V(y) = \mu(1-\mu)/(1+\phi)$を用いて、次式のように置き換えることができる

$$B(\mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma\big((1-\mu)\phi\big)} \cdot y^{\mu\phi-1} \cdot (1-y)^{(1-\mu)\phi-1},$$

$$0 \leq y \leq 1$$

- ここで、

$$\alpha = \mu/\phi$$
$$\beta = (1-\mu)/\phi$$

ベータ回帰モデル

- 被説明変数yが従うベータ関数$y \sim B(\mu, \phi)$

$$B(\mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma\big((1-\mu)\phi\big)} \cdot y^{\mu\phi-1} \cdot (1-y)^{(1-\mu)\phi-1},$$
$$0 \leq y \leq 1$$

- リンク関数はlogitリンク関数

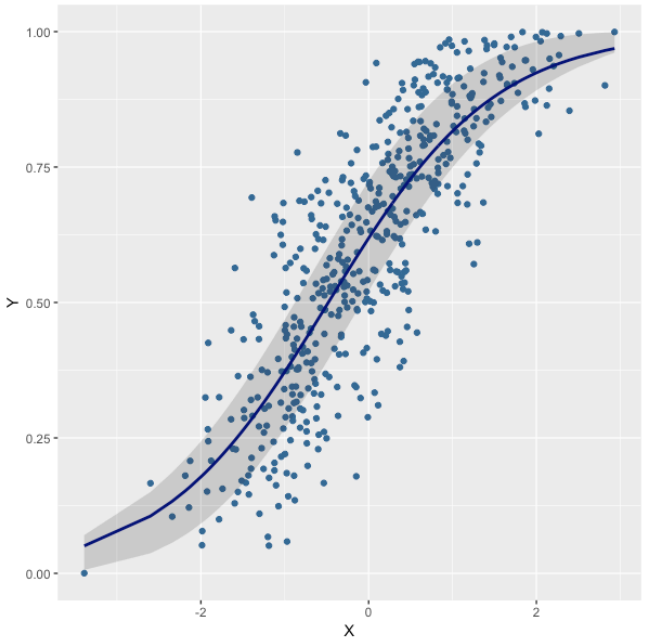
$$g\big(E(y_i)\big) = g\left(E\big(B(\mu_i, \phi_i)\big)\right) = g(\mu_i)$$
$$\text{logit}(\mu_i) = \eta_i$$

- 線形予測子との関係は以下のようになる

$$\mu_i = \frac{1}{1+\exp(-\eta_i)} = \frac{1}{1+\exp\left(-\left(X_i^T \boldsymbol{\beta}\right)\right)}$$

ベータ回帰モデル

- 分析例



ガンマ関数

- ガンマ関数$\Gamma(\alpha)$は次式で表される

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$$
$$\alpha > 0$$

- ガンマ関数$\Gamma(\alpha)$は以下のような性質を持つ

$$\Gamma(\alpha) = (\alpha - 1)!$$
$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$$
$$\Gamma(1/2) = \sqrt{\pi}$$

ガンマ分布

- ガンマ分布の確率密度関数$f(x)$は次式のようになる

$$f(x) = Ga(\alpha, \lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, x \geq 0$$

- ガンマ分布の確率変数をXとするとき、平均$E(X)$と分散$V(X)$は以下の通り（αはshapeパラメータ、$\frac{1}{\lambda}$はscaleパラメータ）

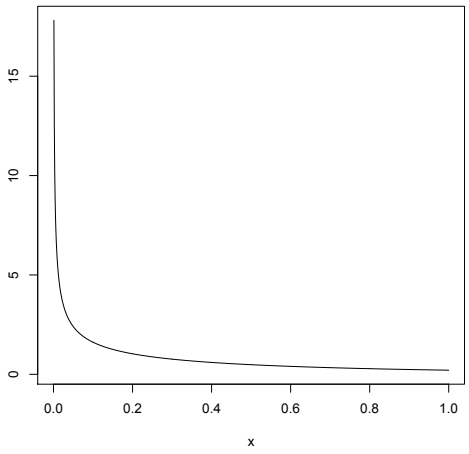
$$E(X) = \frac{\alpha}{\lambda}$$

$$E(X) = \frac{\alpha}{\lambda^2}$$

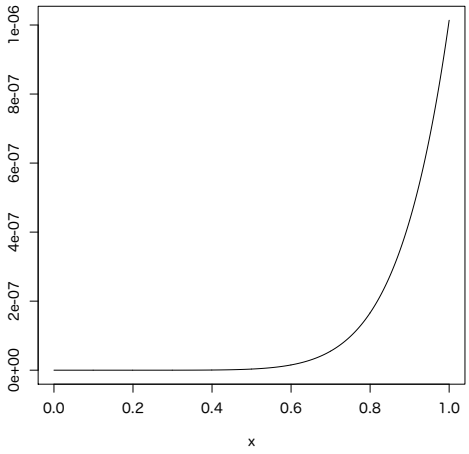
$\frac{1}{\lambda} = \theta$などと記述する場合もある

ガンマ分布

$\Gamma(0.5)$

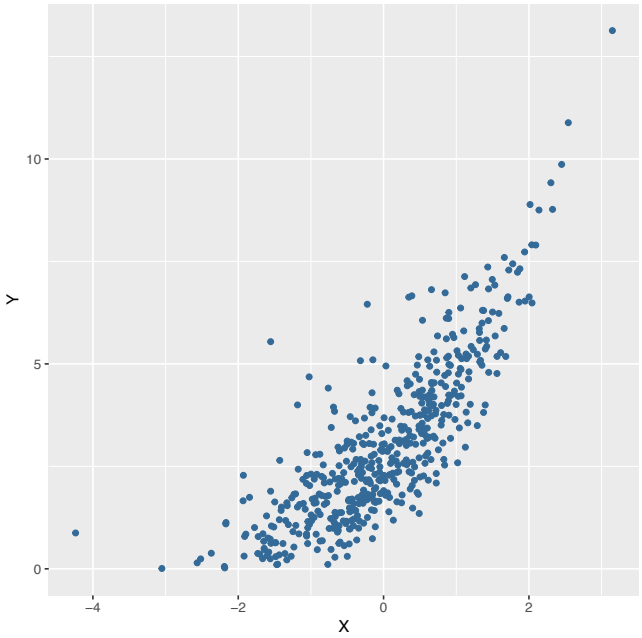


$\Gamma(10)$



ガンマ回帰モデル

- 被説明変数がガンマ分布に従い、非負の実数をとる



ガンマ回帰モデル

- 被説明変数がガンマ分布に従うガンマ回帰モデルは、以下のように表される

$$y_i = Ga(\alpha, \lambda_i)$$

- リンク関数$g\big(E(y_i)\big)$は対数リンク関数の逆関数で表される

$$g\big(E(y_i)\big) = g\left(E\big(Ga(\alpha_i, \lambda_i)\big)\right) = g\left(\frac{\lambda_i}{\alpha_i}\right)$$

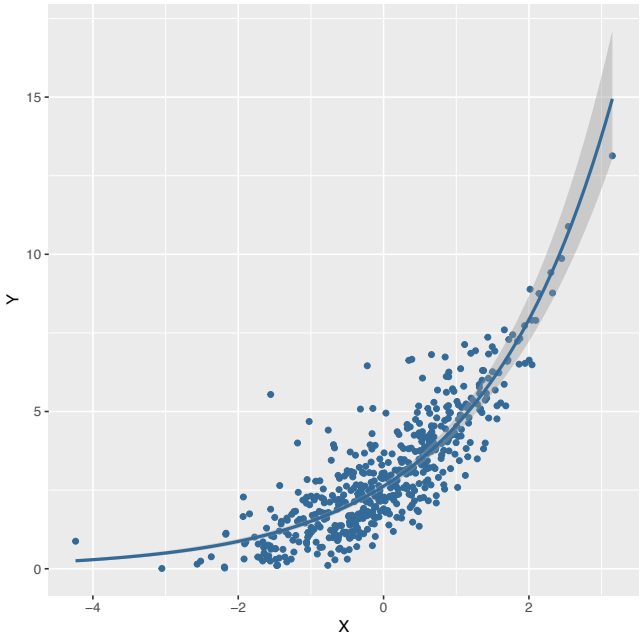
$$\frac{\lambda_i}{\alpha_i} = \frac{1}{\exp(\eta_i)}$$

- logリンクと線形予測子η_iとの関係は以下のように表せる

$$\ln\left(\frac{\alpha_i}{\lambda_i}\right) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = X_i^T \boldsymbol{\beta}$$

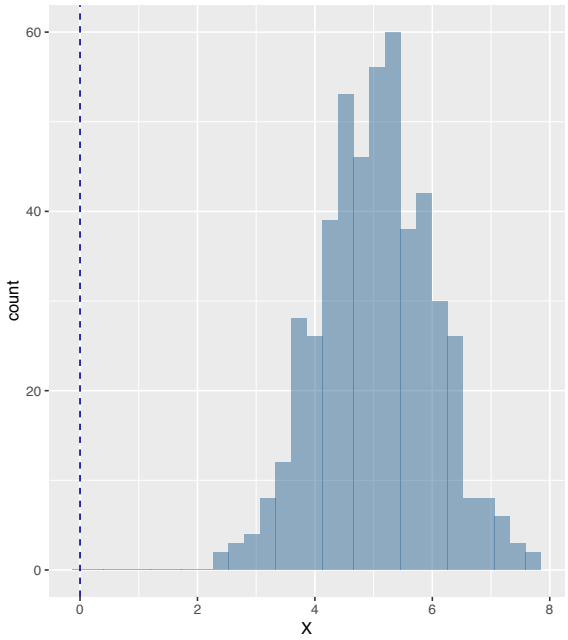
ガンマ回帰モデル

- 分析例



指数-ガウス分布

- 指数分布と正規分布（ガウス分布）の混合分布（指数分布と正規分布の積を畳み込み積分した分布）
- 正規分布の平均μと分散σ^2、指数分布のrateパラメータνの３つのパラメータを持つ
- 反応時間などの分布をモデリングする際に用いられる



指数-ガウス分布

- 指数-ガウス分布は、次式のように表せる

$$y = f(y|\mu, \sigma, \nu)$$
$$= \frac{1}{\nu\sqrt{2\pi}} \exp\left(\frac{\mu - y}{\nu} + \frac{\sigma^2}{2\nu^2}\right) \cdot \int_{-\infty}^{[(y-\mu)/\sigma]-(\sigma/\nu)} \exp\left(-\frac{y^2}{2}\right) dy$$

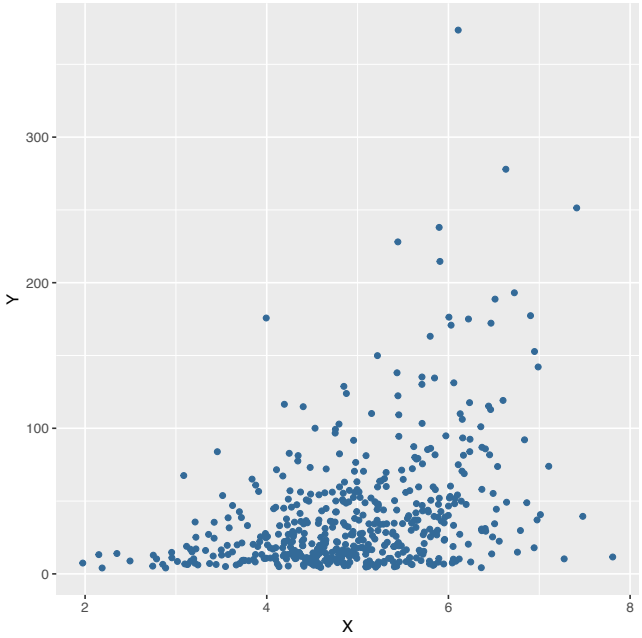
$$-\infty < y < \infty, -\infty < \mu < \infty, \sigma > 0 \text{ and } \nu > 0$$

- 平均と分散は以下のようになる

$$E(y) = \mu + \nu$$
$$V(y) = \sigma^2 + \nu^2$$

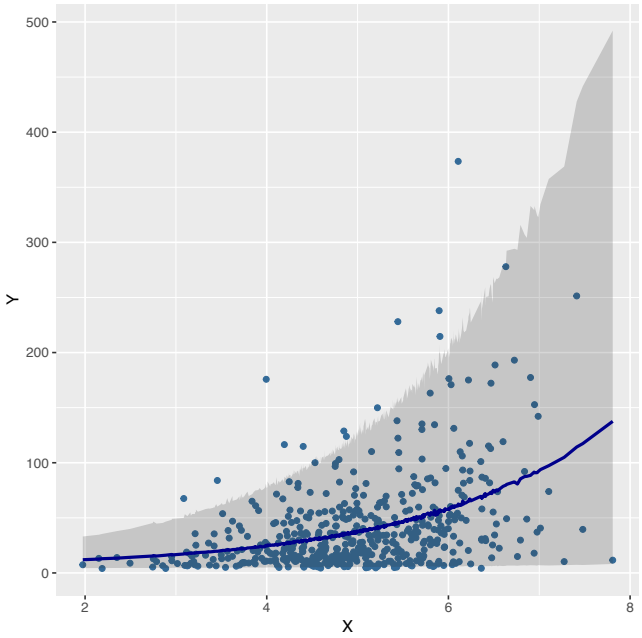
指数-ガウス回帰モデル

- 指数関数のパラメータνに対する回帰モデル
- 指数-ガウス回帰モデルが用いられるデータ分布は、右図のようにXが増加するほど、指数関数の期待値が増加する（右に裾が長い）



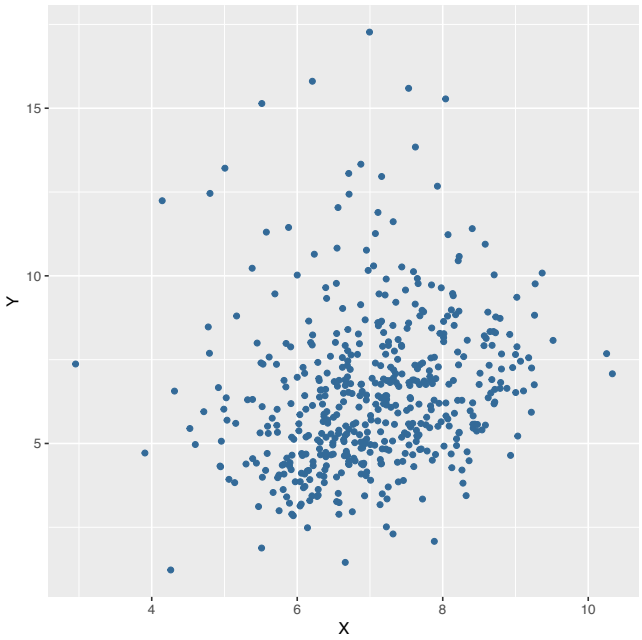
指数-ガウス回帰モデル

- 指数関数のパラメータνに対する回帰モデルの推定結果



指数-ガウス回帰モデル

- 正規分布のパラメータμに対する回帰モデル
- 指数-ガウス回帰モデルが用いられるデータ分布は、右図のようにXが増加するほど、指数関数の期待値が増加する（右に裾が長い）



指数-ガウス回帰モデル

- 正規分布のパラメータμに対する回帰モデルの推定結果