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

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

非負整数を被説明変数とする回帰モデル

- 適用されるモデルの例
  - ポアソン回帰モデル
  - 負の二項分布モデル

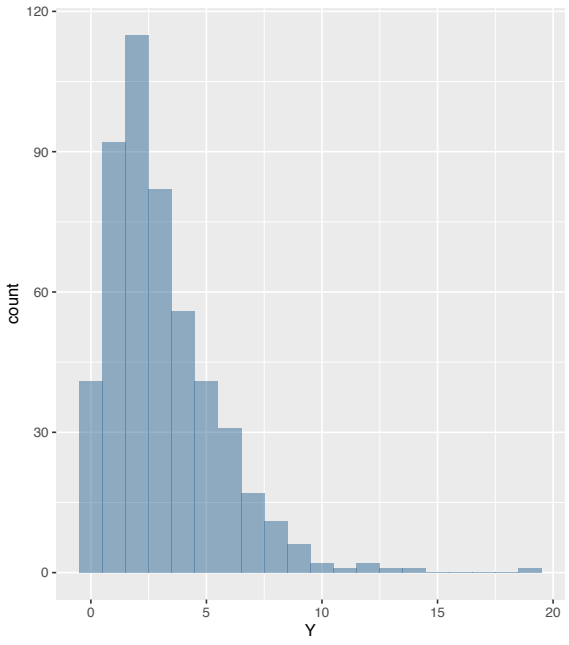
  →被説明変数に 0 が多い場合、ゼロとそれ以上を分けて説明する回帰モデルが用いられることがある（第 6 回授業）
- 適用例
  - 交通事故発生件数、工場で発生する不良品の数、サッカーの得点、放射線のカウント数、単位空間あたりの植物（生物）の数

ポアソン分布

- 非常に多くの観測回数が繰り返されるが、観測ケースの発生頻度が非常に低い場合に用いられる確率分布
- 一定範囲内（時間、回数、空間）である事象が発生する平均=λをもちいて計算される
- 試行回数n、発生頻度pとすると、$\lambda = np$
- ポアソン分布は次式で表される

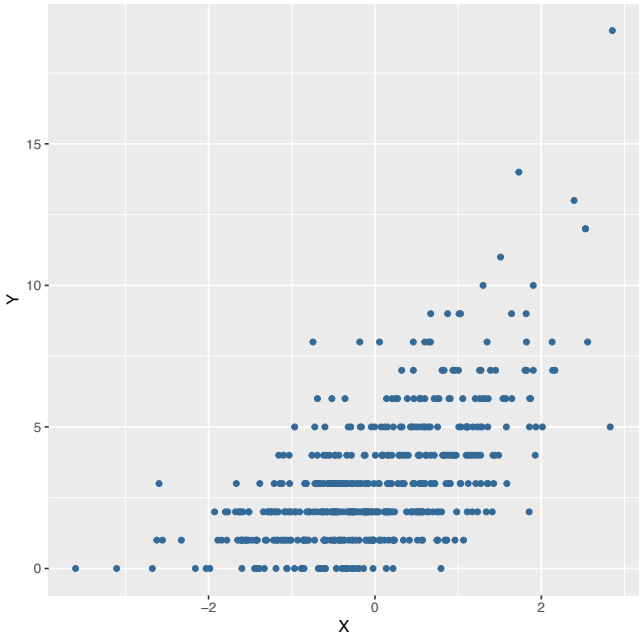
$$P(y|\lambda) = \frac{\lambda^y \cdot \exp(-\lambda)}{y!}$$

- $E(y) = \lambda$、 $Var(y) = \lambda$



ポアソン回帰モデル

- ポアソン回帰モデルは、右図のような関係にあるデータに適用される
- 被説明変数は 0 以上の整数
- 適用事例
  - スポーツの得点分布
  - 感染症の感染者数
  - 事故発生件数
  - イベント来訪者数



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

$$y_i = Po(\lambda_i)$$
$$f(y_i|\lambda_i) = \frac{{\lambda_i}^{y_i} \cdot \exp(-\lambda_i)}{y_i!}$$
$$g\big(E(y_i)\big) = \lambda_i = \exp(\eta_i)$$

- 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

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

ポアソン回帰モデル

- 尤度関数Lは

$$L = \prod_{i=1}^{n} \frac{{\lambda_i}^{y_i} \cdot \exp(-\lambda_i)}{y_i!} = \prod_{i=1}^{n} \frac{\left(\exp(X_i\boldsymbol{\beta})^{y_i}\exp\left(-exp(X_i\boldsymbol{\beta})\right)\right)}{y_i!}$$

- 対数尤度関数$\ln L$は

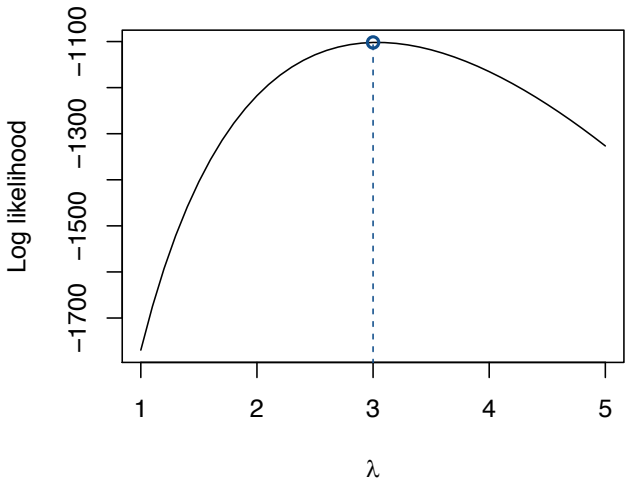
$$\ln L = \sum_{i=1}^{n}\left(y_i \cdot \ln\lambda_i - \lambda_i - \sum_{k=1}^{y_i}\ln k\right)$$
$$= \sum_{i=1}^{n}\left(y_i \cdot \ln\left(X_i^T\boldsymbol{\beta}\right) - \left(X_i^T\boldsymbol{\beta}\right) - \sum_{k=1}^{y_i}\ln k\right)$$

- 対数尤度関数を最大化する$\hat{\lambda}$は

$$\frac{d\ln L}{d\lambda} = 0$$

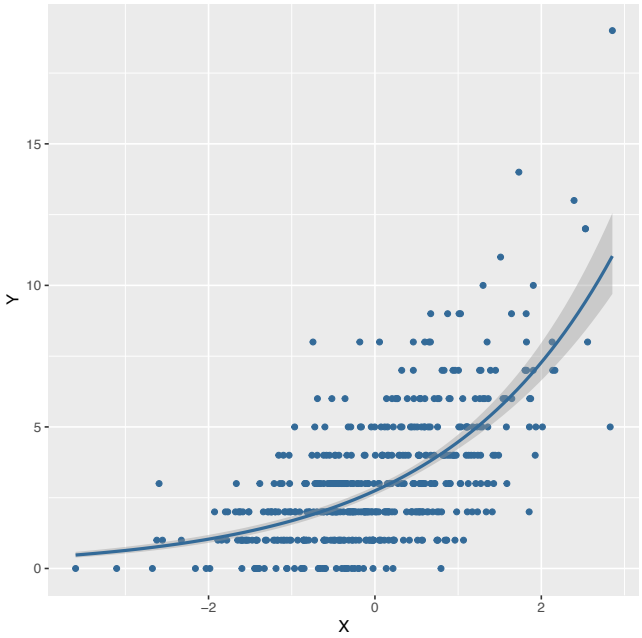
ポアソン回帰モデルの尤度関数

- ポアソン回帰モデルの尤度関数は右図のよう上に凸な関数となる
- この例では、$\lambda = 3.0$が最尤推定値$\hat{\lambda}$となる
- 最尤推定値$\hat{\lambda}$を代入した最大対数尤度は$\ln\hat{L} = -1101.96$



ポアソン回帰モデルの推定結果の例

- 500個のランダムなポアソン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ポアソン回帰モデルを推定
- 得られたモデルの曲線と95%信頼区間は右図のとおり



ポアソン回帰モデルのベイズ推定

- 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_1は説明変数、 (β_0, β_1)は未知パラメータである。

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

- 回帰係数が正規分布に従うので、例えば以下のように事前情報を設定してベイズ推定する。

$$\boldsymbol{\beta} \sim N(0, 1\times10^{-6})$$

ポアソン分布と過分散

- ポアソン分布は、平均が決まると分散が自動的に決まるため、データのばらつきに応じて分散を決めることができない
- 実際のデータの分散が、推定されたモデルを用いて期待される分散との間に乖離がないか確かめる必要がある
- 分散が期待される分散より大きい場合、「過分散」が生じているという
- 過分散の場合、仮定したモデルでは実際のデータより誤差が小さく推定され、z値が大きくなり、有意差が出やすくなる
- つまり、第一種の過誤が生じやすくなる

過分散への対処方法

- 過分散への対処方法として、２つ提案されている
- 疑似ポアソン (quasi-Poisson) モデルを用いる方法
- 負の二項分布モデルを用いる方法

疑似ポアソン回帰モデル

- モデルから期待される分散に比べて、実際のデータの分散がどの程度大きいかを計算し、そのばらつきの度合い (dispersion parameter) θに合わせてモデルの誤差を調整したモデル
- ポアソン分布の期待値$E(y) = \lambda$に対して、dispersion parameterを考慮した分散は$Var(y) = \theta\lambda$となる
- Dispersion parameterθは残差の自由度$df.res$を用いて、次式から得られる

$$\theta = \frac{\chi^2}{df.res} = \frac{\sum_{i=1}^{n} \left(y_i - \hat{\lambda}_i\right)^2 / \hat{\lambda}_i}{df.res}$$

負の二項分布

- 過分散があると考えられるデータを表す確率分布の一つに、負の二項分布がある
- 負の二項分布は以下のように表せる

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

- これは、ポアソン分布のλがガンマ分布$Ga(\phi^{-1}, \lambda\phi)$に従うとした確率分布関数である

負の二項分布

- 負の二項分布

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

- このとき

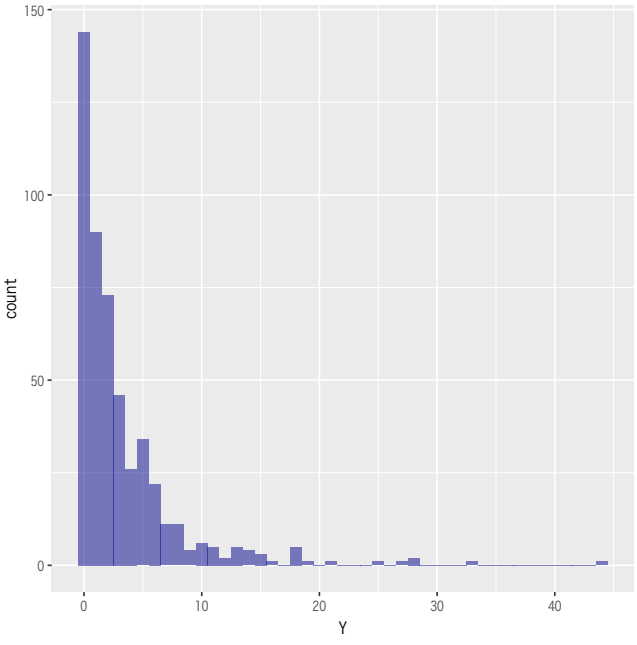
平均： λ

分散： $\frac{\lambda^2}{\phi}+\lambda$

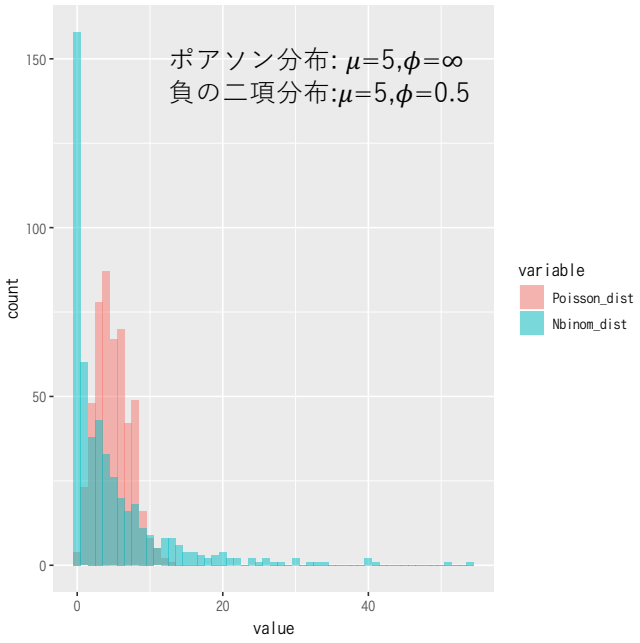
- $(1+\phi)/\phi$を分散指標、 ϕをscale parameterという
- $\phi \to +\infty$としたとき、負の二項分布はポアソン分布となる

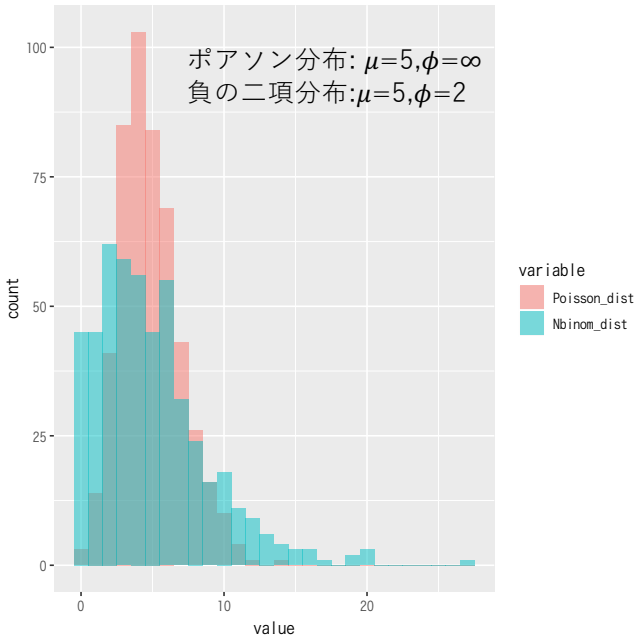
負の二項分布

- 負の二項分布モデルは、非負の実数をとる
- 試行回数、成功確率または分布平均を指定する
- ポアソン分布と比較して分散が大きい
- 「０」値のカウント数が大きい場合もある



ポアソン分布と負の二項分布





二項分布と負の二項分布

- 二項分布： n回の試行でr回成功（ベルヌーイ試行）

$$P(y) = {}_nC_r \cdot \pi^r \cdot (1-\pi)^{n-r}$$

- 負の二項分布：ある事象についてr回までその事象が生じなかった確率分布

$$P(y) = {}_{n-1}C_{r-1} \cdot \pi^r \cdot (1-\pi)^{n-r}$$

- 過分散が生じるような現象に対して用いられる

二項分布と負の二項分布

- ここで$n = y + r$と置き換えると、負の二項分布はガンマ関数を用いて以下のように表せる

$$P(y) = {}_{y+r-1}C_{r-1} \cdot \pi^r \cdot (1-\pi)^y$$
$$= \frac{(y+r-1)!}{y!\,(r-1)!} \cdot \pi^r \cdot (1-\pi)^y$$
$$= \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} \cdot \pi^r \cdot (1-\pi)^y$$

- このとき

$$\text{平均：}\ \frac{r(1-\pi)}{\pi}$$
$$\text{分散：}\ \frac{r(1-\pi)}{\pi^2}$$

二項分布と負の二項分布

- さらに$\pi = \phi/(\lambda + \phi)$と置き換えると、負の二項分布は以下のように表せる

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

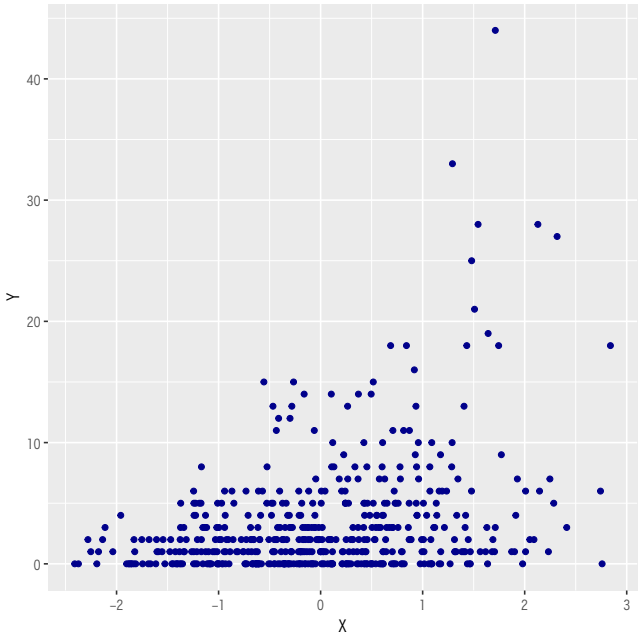
- このとき

$$\text{平均：}\ \lambda$$
$$\text{分散：}\ (\lambda + \phi\lambda)/\phi$$

- $(1+\phi)/\phi$を分散指標、 ϕをscale parameterという

負の二項分布モデル

- ポアソン分布と比較して過分散と考えられる場合や、0値のカウント数が多いと考えられる場合に用いられる



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

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

負の二項分布モデル

- 尤度関数Lは

$$L = \prod_{i=1}^{n} \frac{\Gamma(y_i + \phi)}{\Gamma(y_i + 1)\Gamma(\phi)} \left(\frac{\phi}{\phi + \lambda_i}\right)^{\phi} \left(\frac{\lambda_i}{\phi + \lambda_i}\right)^{y_i}$$

- 対数尤度関数$\ln L$は

$$\begin{aligned}
&\ln L_n \\
&= \sum_{i=1} \{\ln[\Gamma(y_i + \phi)] - \ln[\Gamma(y_i + 1)] - \ln[\Gamma(\phi)] + \phi \ln \phi \\
&- \phi \ln(\phi + \lambda_i) + y_i \ln \lambda_i - y_i \ln(\phi + \lambda_i)\} \\
&= \sum_{i=1}^{n} \{\ln[\Gamma(y_i + \phi)] - \ln[\Gamma(y_i + 1)] - \ln[\Gamma(\phi)] + \phi \ln \phi + y_i \ln \lambda_i \\
&- (\phi + y_i) \ln(\phi + \lambda_i)\}
\end{aligned}$$

負の二項分布モデル

- 対数尤度関数を最大化する$\hat{\beta}_j$および$\hat{\lambda}$は

$$\frac{d \ln L}{d \beta_j} = \sum_{i=1}^{n} \frac{x_{ij}(y_i - \lambda_i)}{1 + \lambda_i / \phi} = 0$$

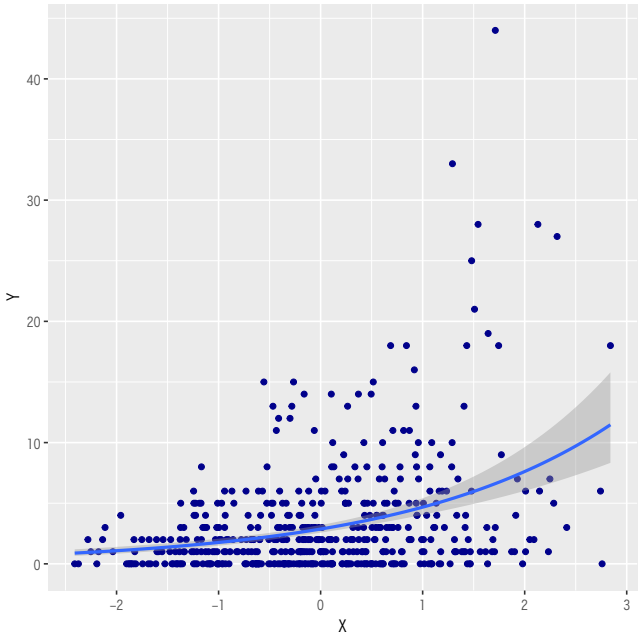
$$\frac{d \ln L}{d \lambda} = \sum_{i=1}^{n} \left\{ \phi^2 \left(\ln(1 + \lambda_i / \phi) - \sum_{j=0}^{y_i - 1} \frac{1}{j + \phi} \right) + \frac{\phi(y_i - \lambda_i)}{1 + \lambda_i / \phi} \right\} = 0$$

- ここで、ガンマ関数は以下の性質を持つ

$$\ln\left(\frac{\Gamma(y_i + \phi)}{\Gamma(\phi)}\right) = \sum_{j=0}^{y_i - 1} \frac{1}{j + \phi}$$

負の二項分布モデルの推定結果の例

- 500個のランダムな負の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負の二項分布回帰モデルを推定
- 得られたモデルの曲線と95%信頼区間は右図のとおり



負の二項分布モデルのベイズ推定

- 負の二項分布モデル

$$y_i \sim NB(r, \pi_i)$$

$$\pi_i = \frac{r}{\lambda_i + r}$$

$$log(\lambda_i) = \beta_0 + \beta_1 x_{1i}$$

- 事前情報は例えば以下のようにする

$$\beta_0, \beta_1 \sim N(0, 1\times10^{-2})$$

$$r \sim Ga(1\times10^{-3}, 1\times10^{-3})$$