NagoyaStat#3
データ解析のための統計モデリング入門6章

@nonsabotage
2016.11.26 ヤフー株式会社名古屋オフィス

自己紹介

- @nonsabotage
- 建設コンサルタント
- 緑本は3年振り8周目（そろそろ、理解したいが…

6

GLMの応用範囲をひろげる

- ロジスティック回帰など -

6.1 さまざまなデータで応用できるGLM

6章の目的:

GLMの3要素に関する技術を学び、

GLMでモデリングできるデータを増やす

- 応答変数の確率分布
- 線形予測子
- リンク関数(応答変数の平均と線形予測子の関係)

6.2 例題：上限のあるカウントデータ

6.3〜6.5へ向けてのデータの説明

観測データ:

「ある植物 i において8個 の観察種子の発芽能力があるものは y_i 個、死んだ種子は$8 - y_i$ 個」

データサイズ:

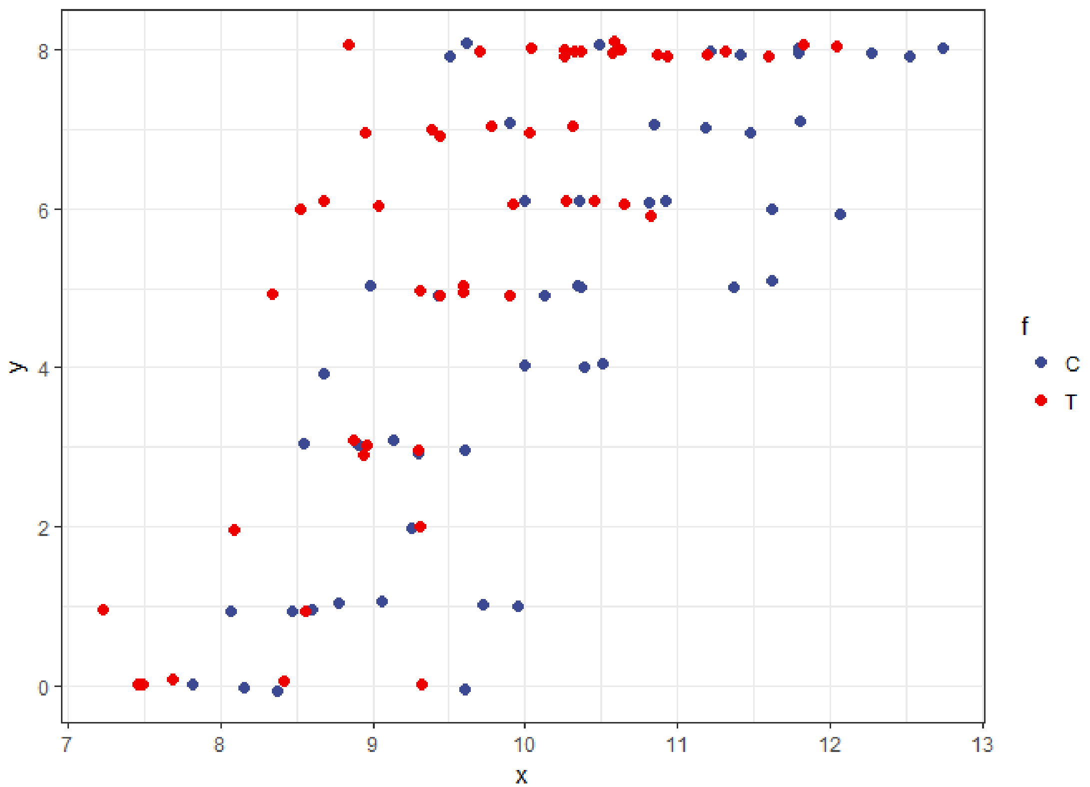
100 個体

分析の目的:

生存確率q_iに体サイズx_iや施肥処理f_iが与える影響

6.2 例題：上限のあるカウントデータ

データの散布図



Xが増えると
Yも増えそう

赤い点の方が
Yが大きそう

6.3 二項分布で表現するカウントデータ

応答変数の確率分布を設定する

$$y \in \{0, 1, ...\}$$

ポアソン分布では上限のない
カウントデータは表現できた

$$y \in \{0, 1, ..., 8\}$$

上限があるカウントデータを
表現する確率分布は？

⇒ **二項分布**

6.3 二項分布で表現するカウントデータ

二項分布について

結果が成功・失敗で評価できる試行を、成功率qのもとで独立にN回試行した際の成功数の離散確率分布

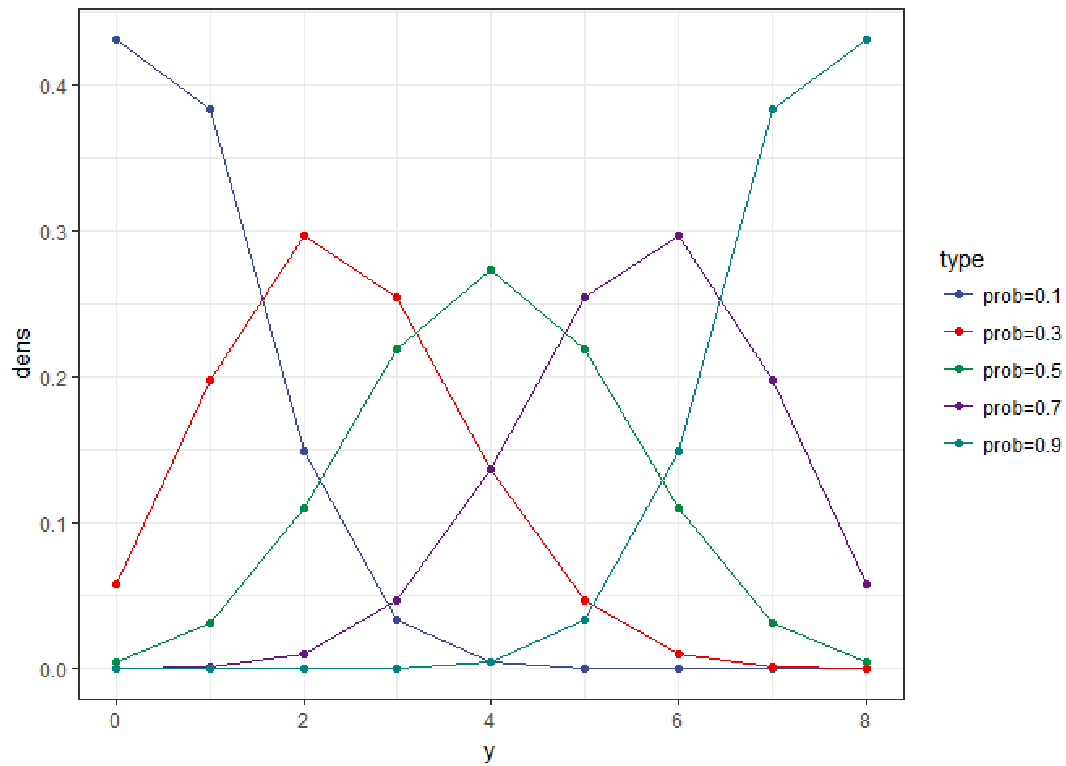
$$p(y|N,q) = \binom{N}{y} q^y (1-q)^{N-y}$$

$$E[y] = Nq$$

$$V[y] = Nq(1-q)$$

6.3 二項分布で表現するカウントデータ

二項分布の確率密度関数



6.4 ロジスティック回帰とロジットリンク関数

リンク関数の設定

$$E[y] = Nq$$

qをモデリングす(Nは決まっている)

$$q \in (0, 1)$$

qは0~1の実数

$$g(q) = \sum_i \beta_i x_i$$

リンク関数gはなにを使えば？

⇒ **ロジットリンク関数**

6.4 ロジスティック回帰とロジットリンク関数

ロジットリンク関数について

$$\mathrm{logit}(q_i) = \log \frac{q_i}{1-q_i}$$

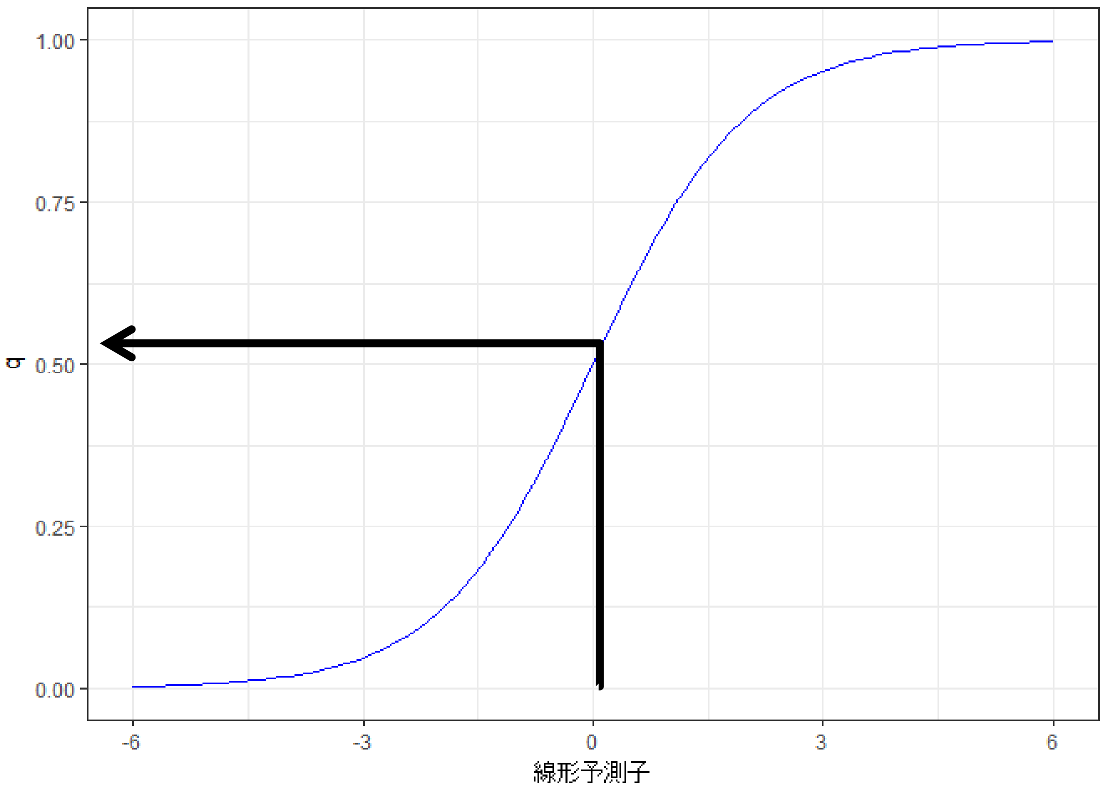
ロジットリンク関数の定義

$$q_i = \mathrm{logit}^{-1}\left(\sum_i \beta_i x_i\right)$$
$$= \frac{1}{1+\exp(-\sum_i \beta_i x_i)}$$

ロジットリンク関数の逆関数が
ロジスティック関数となる

6.4 ロジスティック回帰とロジットリンク関数

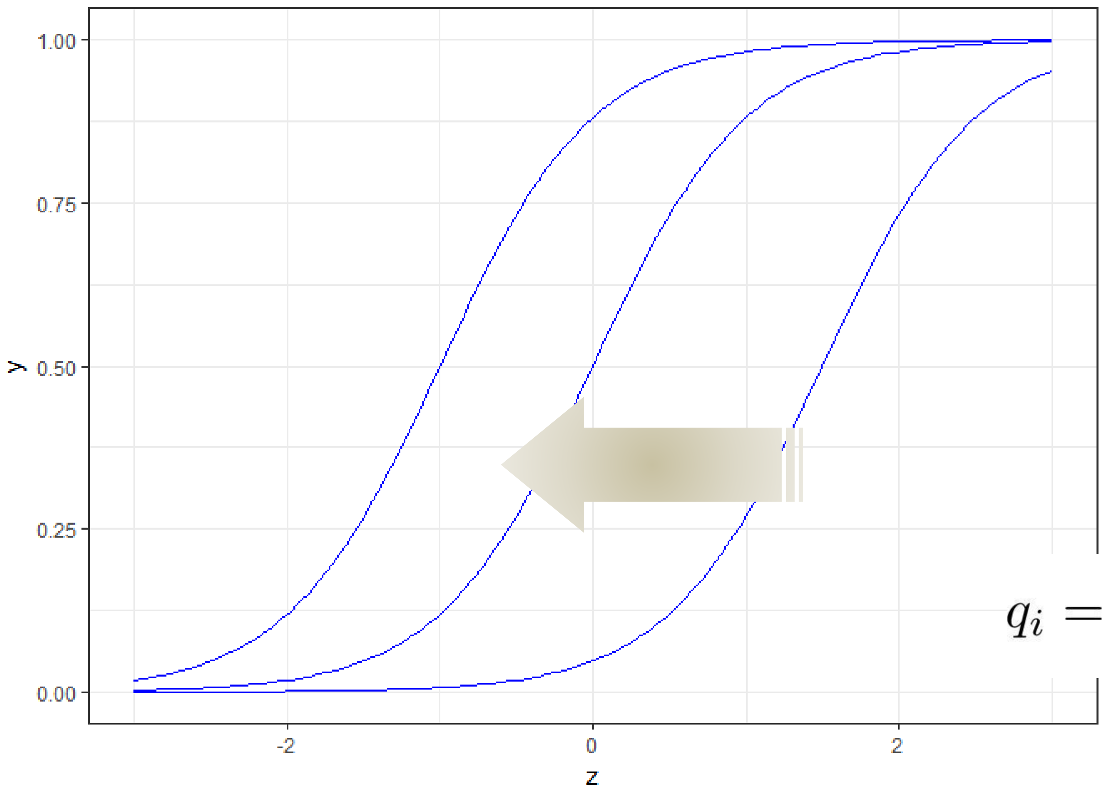
ロジスティック関数について



線形予測子を
0~1へ写像

6.4 ロジスティック回帰とロジットリンク関数

ロジスティック関数でβ_0が増加すると

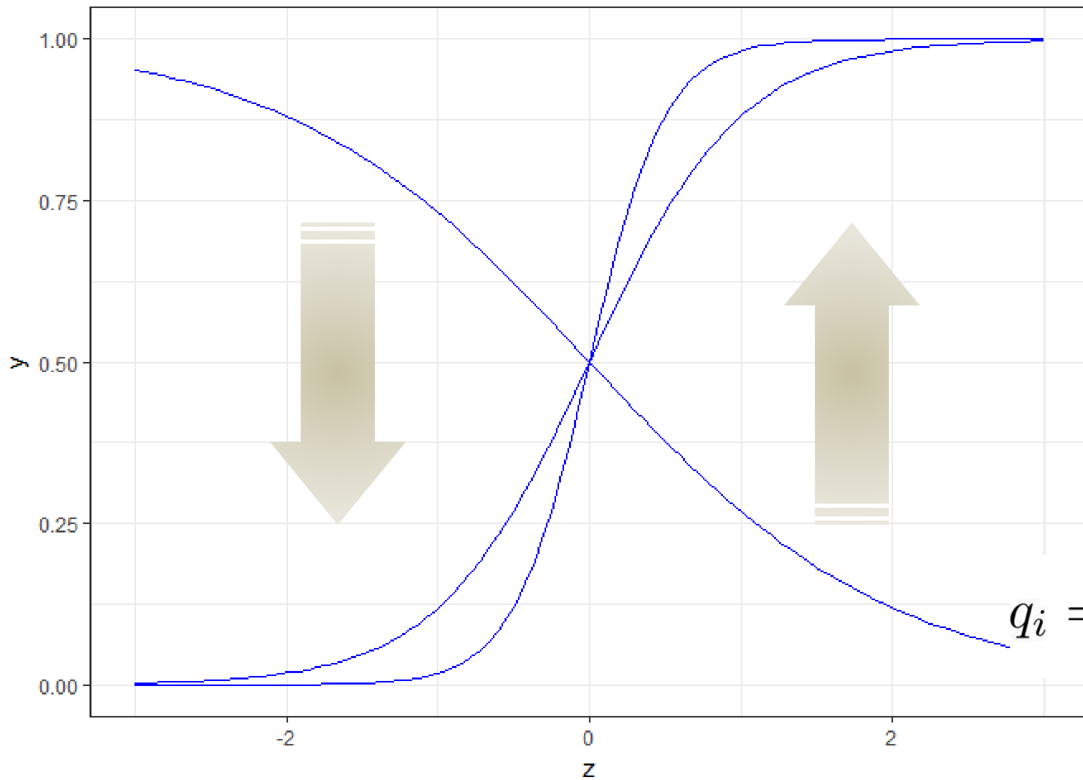


定数項が大きいほど左へシフト

$$q_i = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x_1)\}}$$

6.4 ロジスティック回帰とロジットリンク関数

ロジスティック関数でβ_1が増加すると



変化が急激に

$$q_i = \frac{1}{1+\exp\{-(\beta_0+\beta_1 x_1)\}}$$

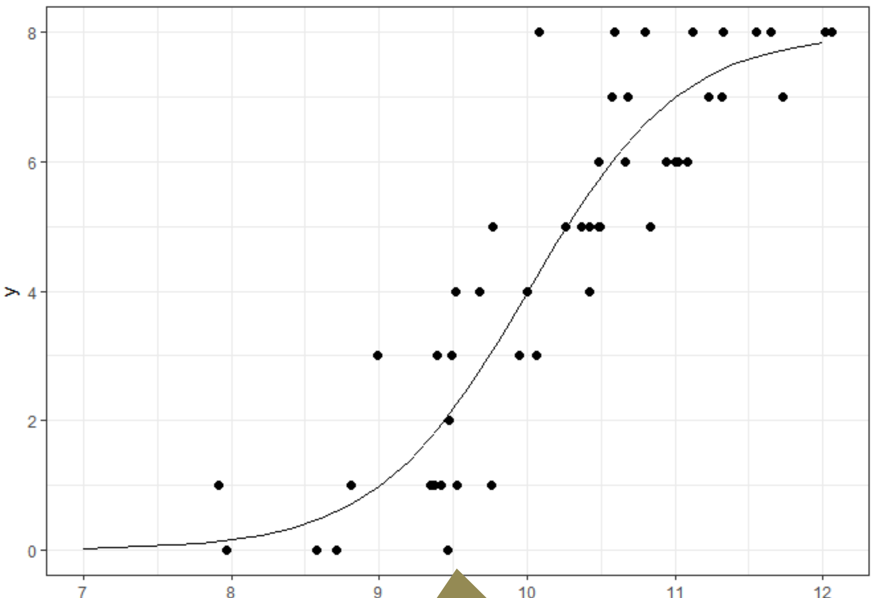
6.4 ロジスティック回帰とロジットリンク関数

Rによる推定

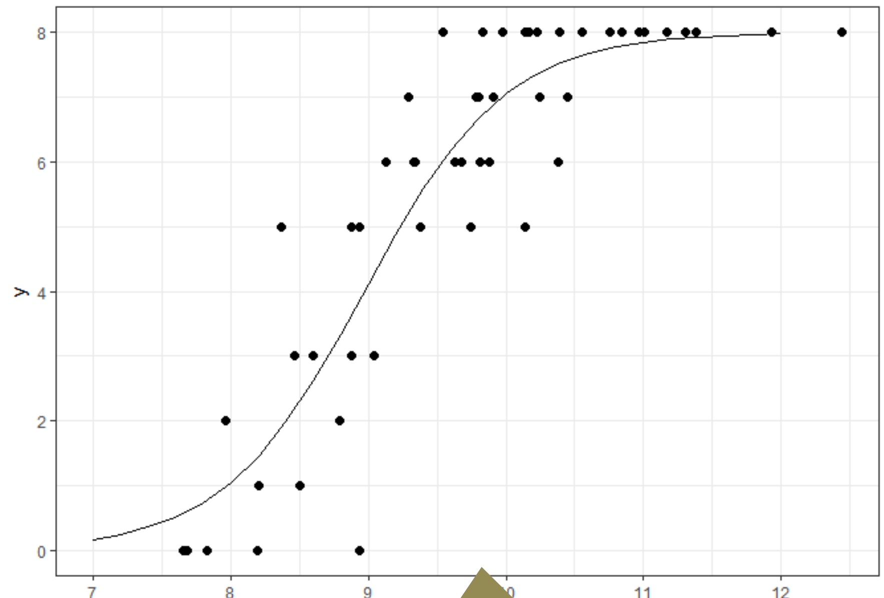
```
> fit <- glm (cbind(y, N-y) ~ x + f, data = obs, family = binomial)
>
> fit

Call:  glm(formula = cbind(y, N - y) ~ x + f, family = binomial, data = obs)

Coefficients:
(Intercept)            x           fT
    -19.536        1.952        2.022

Degrees of Freedom: 99 Total (i.e. Null);  97 Residual
Null Deviance:      499.2
Residual Deviance: 123  AIC: 272.2
>
```

6.4 ロジスティック回帰とロジットリンク関数

推定結果



施肥なし



施肥あり

6.4 ロジスティック回帰とロジットリンク関数

ロジスティック回帰で推定したパラメータの解釈

$$\frac{q_i}{1 - q_i} = \exp(\beta_0)\exp(\beta_1 x_1)\exp(\beta_2 f_i)$$

説明変数が1単位変化した際の
オッズ（左辺）の倍率となる

6.4 ロジスティック回帰とロジットリンク関数

オッズ比とリスク

ロジスティック回帰でパラメータ推定しておくと、

オッズの比で、リスクが近似できる。

「xxな人たちは、そうでない人に比べて〇〇倍 ～しやすい」

$$\frac{q_i}{1-q_i} / \frac{p_i}{1-pi} = \exp(\beta_x)$$

比を取ることで
共通部分が落ちる

6.4 ロジスティック回帰とロジットリンク関数

ロジスティック回帰のモデル選択

name	formula	null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
f	cbind(y, N-y) ~ f	499.2321	99	-316.87988	637.7598	642.9701	490.5825	98
fixed	cbind(y, N-y) ~ 1	499.2321	99	-321.20467	644.4093	647.0145	499.2321	99
x	cbind(y, N-y) ~ x	499.2321	99	-180.17272	364.3454	369.5558	217.1682	98
x*f	cbind(y, N-y) ~ x+f+x:f	499.2321	99	-132.80530	273.6106	284.0313	122.4334	96
x+f	cbind(y, N-y) ~ x+f	499.2321	99	-133.10556	272.2111	280.0266	123.0339	97

x+fがAIC最小

6.5 交互作用項の入った線形予測子

交互作用項はむやみに入れない

name	formula	null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
f	cbind(y, N-y) ~ f	499.2321	99	-316.87988	637.7598	642.9701	490.5825	98
fixed	cbind(y, N-y) ~ 1	499.2321	99	-321.20467	644.4093	647.0145	499.2321	99
x	cbind(y, N-y) ~ x	499.2321	99	-180.17272	364.3454	369.5558	217.1682	98
x*f	cbind(y, N-y) ~ x+f+x:f	499.2321	99	-132.80530	273.6106	284.0313	122.4334	96
x+f	cbind(y, N-y) ~ x+f	499.2321	99	-133.10556	272.2111	280.0266	123.0339	97

組み合わせ爆発

結果が
解釈しずらい

複雑なモデルが
選ばれやすくなる

6.6 割算値の統計モデリングはやめよう

オフセット項を用いた割算値の回避

割算値はなぜいけない？

情報の損失　：3/10と30/100は同じ3割か？

分布の複雑化：（確率変数）/（確率変数）はどんな分布に

割算値をさけるには？

⇒ **オフセット項の導入**

6.6 割算値の統計モデリングはやめよう

例題データの説明

観測データ:

「調査地i ごとに面積A_iが異なる箇所で、
植物の 発見個体数 y_i を記録した」

データサイズ:

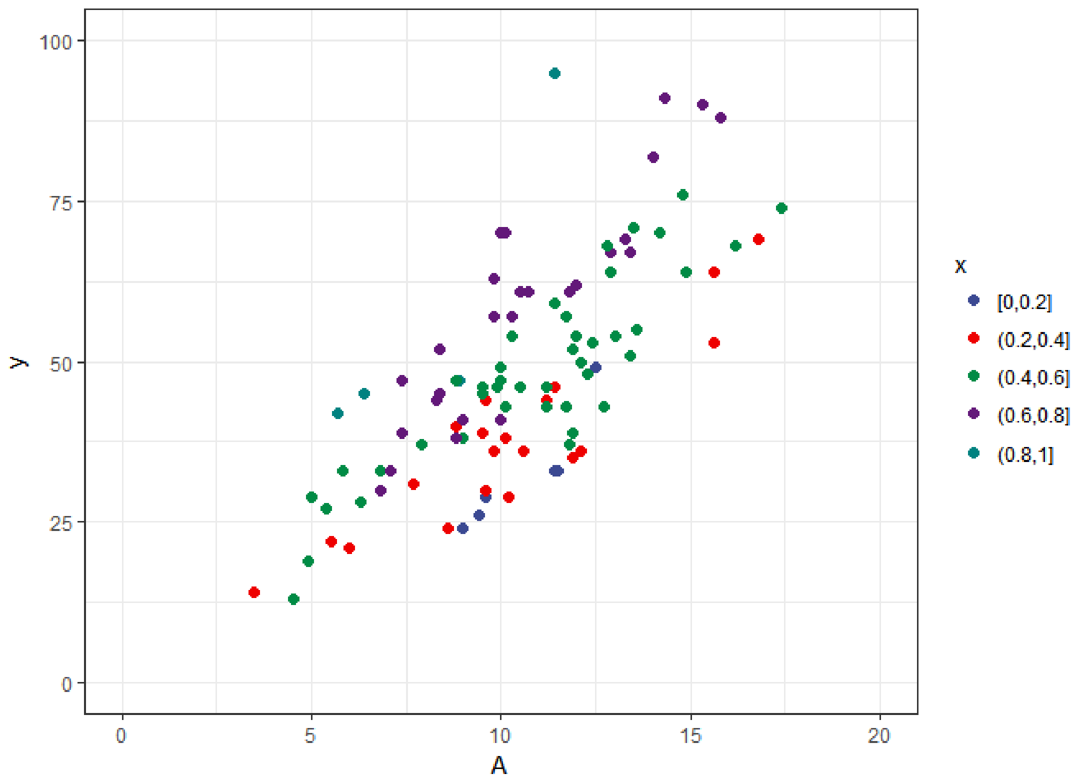
100 箇所

分析の目的:

植物の個体数y_iを箇所iの明るさx_iでモデル化

6.6 割算値の統計モデリングはやめよう

データの散布図



調査面積に比例

明るいほど
傾きが大きそう

6.6 割算値の統計モデリングはやめよう

応答変数の分布を設定する

$\frac{y_i}{A_i}$ が応答変数ではなく、応答変数はy_iでいく。

y_iが上限のないカウントデータなので

応答変数の分布はポアソン分布とする

6.6 割算値の統計モデリングはやめよう

線形予測子を設定する

発見個体数と調査面積とは比例していると考える

比例定数、つまり密度を明るさでモデリングする

$$\lambda_i = A_i \times \text{密度} = A_i \exp(\beta_0 + \beta_1 x_1)$$

$$\log \lambda_i = \log A_i + \beta_0 + \beta_1 x_1$$

ベータがついていない
オフセット項が出現

6.6 割算値の統計モデリングはやめよう

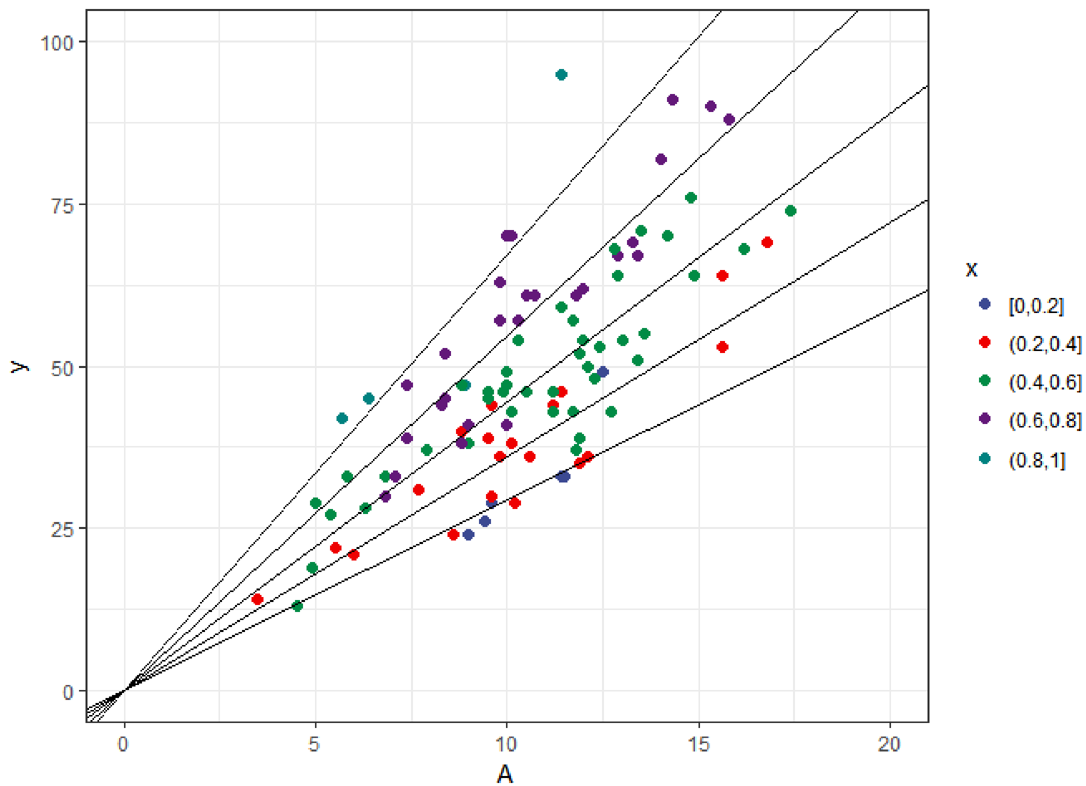
Rによる推定

```
> fit <- glm (y ~ x, offset = log(A), family=poisson, data=obs)
> fit

Call:  glm(formula = y ~ x, family = poisson, data = obs, offset = log(A))

Coefficients:
(Intercept)            x
     0.9731       1.0383

Degrees of Freedom: 99 Total (i.e. Null)
Null Deviance:       261.5
Residual Deviance: 81.61        AIC: 650.3
```

Aをオフセット項とした
推定ができてる

6.6 割算値の統計モデリングはやめよう

推定結果



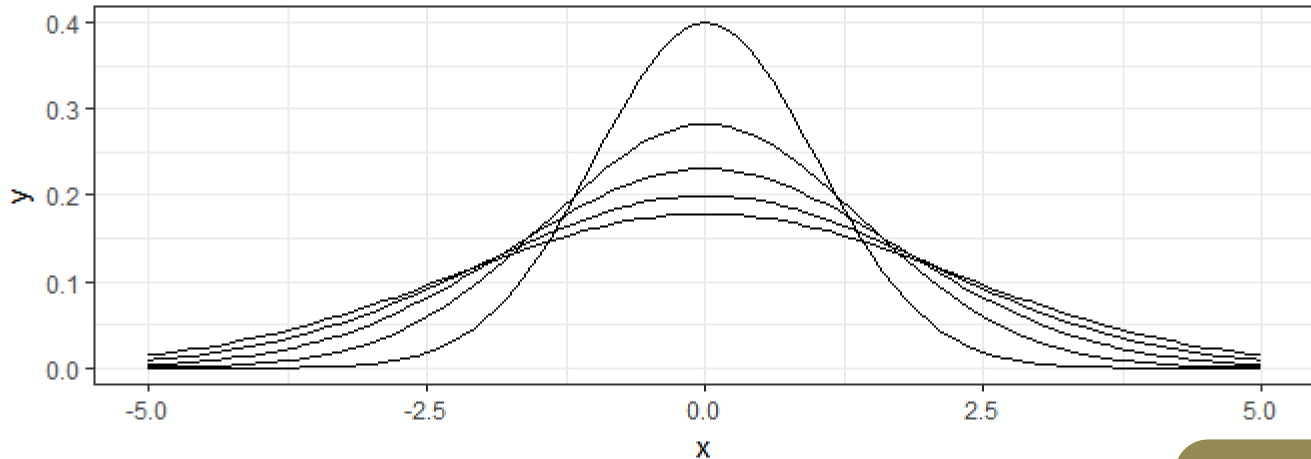
6.7 正規分布とその尤度

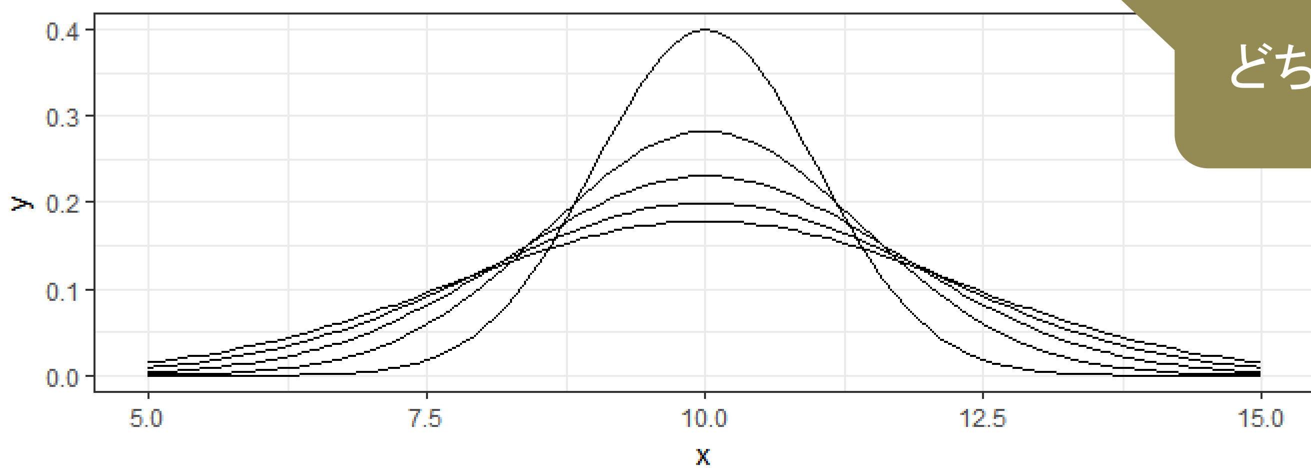
ポイントは3つ

- (-∞, ∞)を取る連続値の確率変数
- 尤度を確率密度関数で考える（正の対数尤度が有り得る）
- 分散一定で、最尤推定値と最小二乗推定値が一致

$$\log L(\mu, \sigma) = -0.5N \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (y_i - \mu)^2$$

6.7 正規分布とその尤度





どちらも分散は1,2,..,5

6.8 ガンマ分布GLM

例題データの説明

観測データ:

「個体i ごとに花重量 y_i を記録した」

データサイズ:

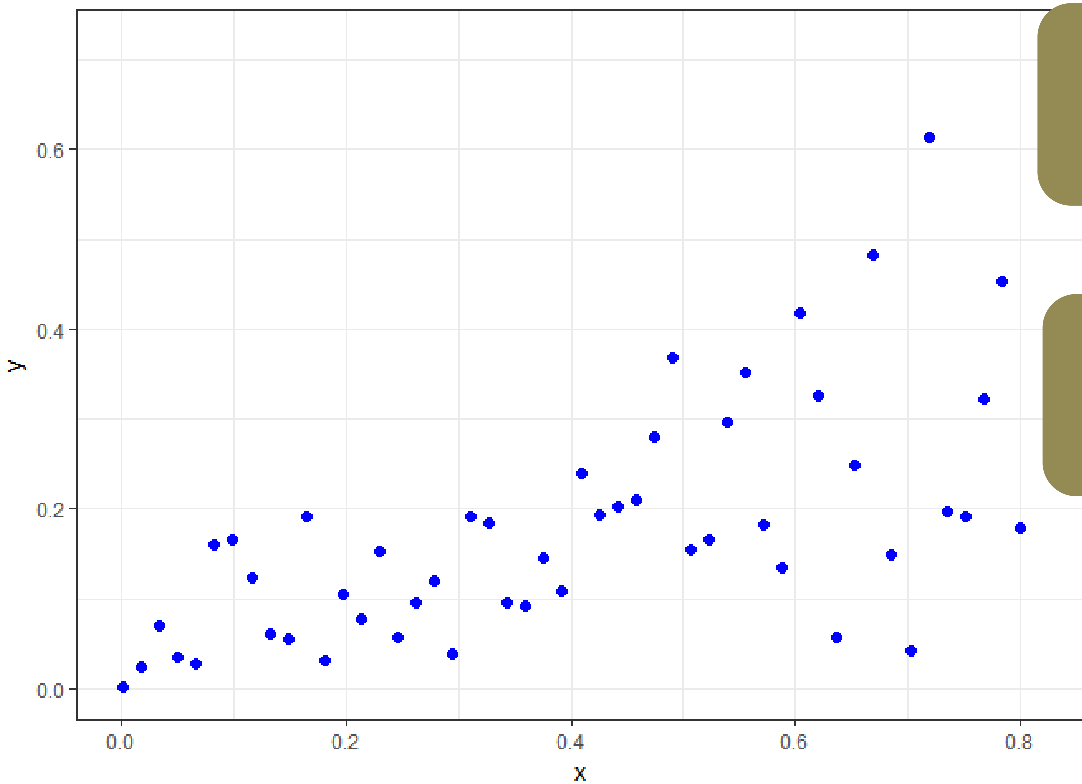
50個体

分析の目的:

花重量y_iを葉重量x_iでモデル化

6.8 ガンマ分布GLM

データの散布図



Xが大きいと
Yも大きい

Xが大きいと
バラつきが大きい

6.8 ガンマ分布GLM

応答変数の確率分布を設定する

$$y \in \{0, 1, ...\}$$

ポアソン分布では上限のない
カウントデータが表現できた

$$y \in [0, \infty)$$

上限のない連続データを
表現する確率分布は？

⇒ **ガンマ分布**

6.8 ガンマ分布GLM

ガンマ分布について

２つのパラメータをもつ、0~∞の値をとる連続確率分布

電子製品の寿命分布などに応用される（wikiより）

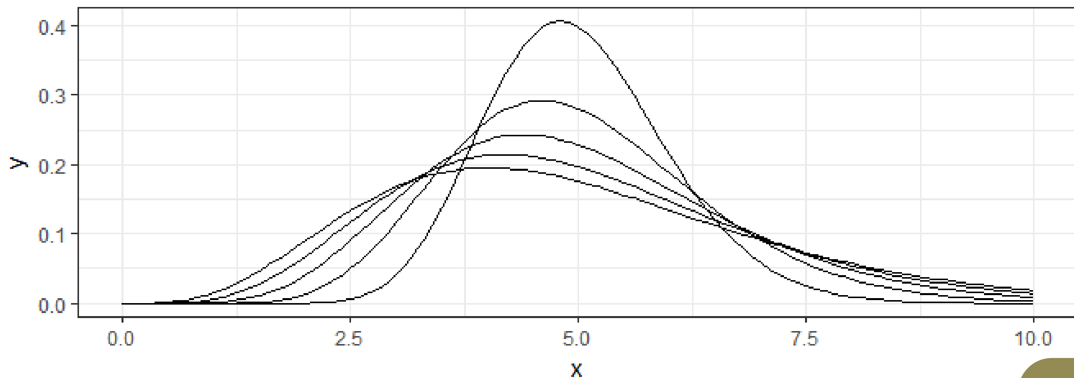
$$f(x|\alpha,\beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x)$$

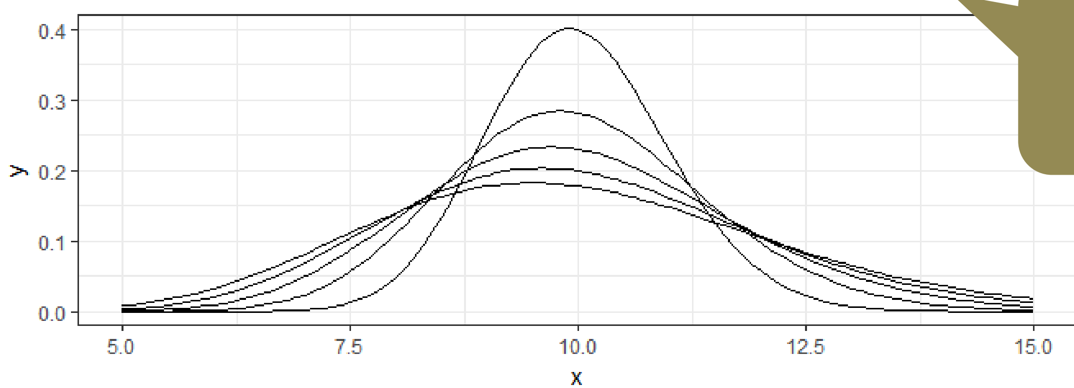
$$E[x]=\frac{\alpha}{\beta}$$

$$V[x]=\frac{\alpha}{\beta^2}$$

6.8 ガンマ分布GLM

ガンマ分布について





どちらも分散は1,2,..,5

6.8 ガンマ分布GLM

線形予測子を設定する

生物学的知識背景から次を設定

$$\mu_i = A x_i^b$$

$$\begin{aligned}\log \mu_i &= \log A + b \log x_i \\ &= a + b \log x_i\end{aligned}$$

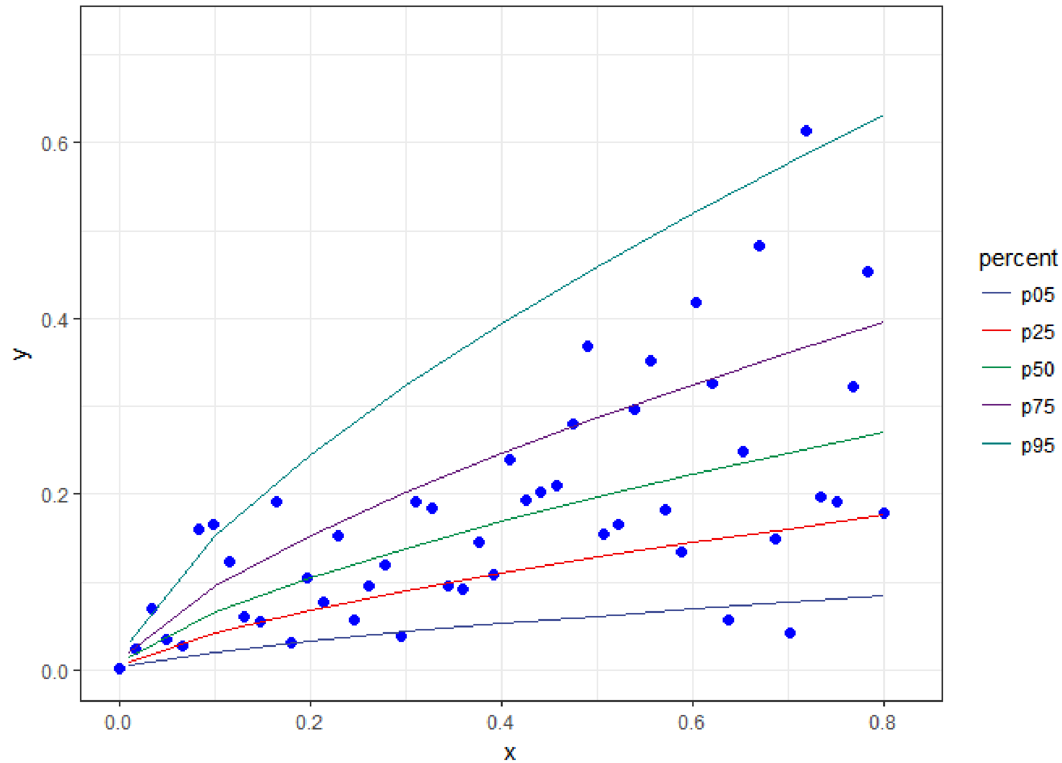
6.8 ガンマ分布GLM

Rによる推定

```
> glmfit      <- glm (y ~ log(x), family = Gamma(link="log"), data=obs)
> glmfit

Call:  glm(formula = y ~ log(x), family = Gamma(link = "log"), data = obs)

Coefficients:
(Intercept)       log(x)
    -1.0403       0.6833

Degrees of Freedom: 49 Total (i.e. Null);  48 Residual
Null Deviance:       35.37
Residual Deviance: 17.25        AIC: -110.9
```

6.8 ガンマ分布GLM

推定結果



6.X まとめ

GLMでモデリングできるデータを増やすための技術を学んだ

- 応答変数の確率分布
  - **二項分布、(正規分布)、ガンマ分布**
- 線形予測子
  - **交互作用項、オフセット項**
- リンク関数(応答変数の平均と線形予測子の関係)
  - **ロジットリンク関数**

おわり