「生態学基礎論 (生物多様性論 II)」の一部:

生態学の統計モデリング (2009 年 1 月) の投影資料

全部で 2 回講義の 1 回目

一般化線形モデル (GLM) の基礎

観測データの統計モデル化

久保拓弥 kubo@ees.hokudai.ac.jp

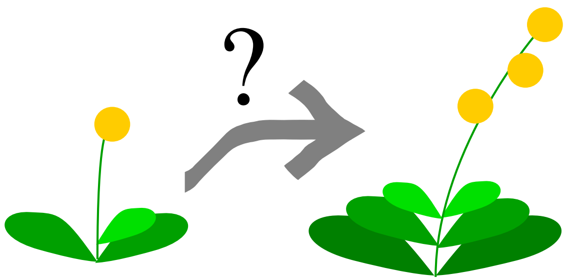
http://hosho.ees.hokudai.ac.jp/~kubo/stat/

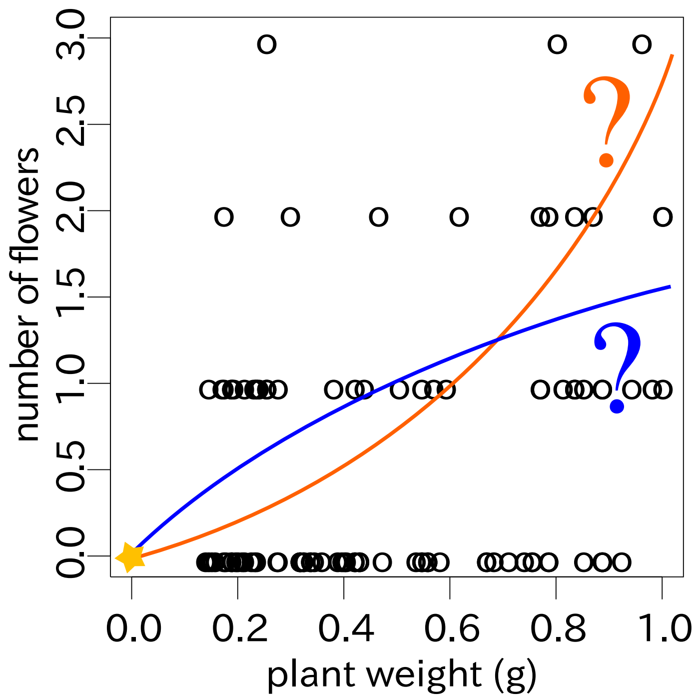
(講義の一とも参照してください)

今日のハナシ

1. この授業の概要: 統計モデルって何なの?
2. 統計ソフトウェア R の簡単な紹介
3. 一般化線形モデル: ポアソン回帰

地上部の重量 $\boldsymbol{x}$
が増加するにつれて
花数 $\boldsymbol{y}$ は増加する
だろうか?





1. この授業の概要:

統計モデルって何なの?

全 2 回だけの授業: 統計モデリングの概要

主題: 一般化線形モデル (GLM) を使った統計モデリングと「脱」割算解析

1. 観測データの統計モデル化 (1/26 月)
   - 統計モデルとは? GLM とは?
   - (GLM の一部である) ポアソン回帰の説明
2. 何でも「割算」するな! (1/28 水)
   - ポアソン回帰を強化する `offset` 項わざ
   - (GLM の一部である) ロジスティック回帰の説明

これらの授業によって何が身につくのか?

この授業の基本姿勢

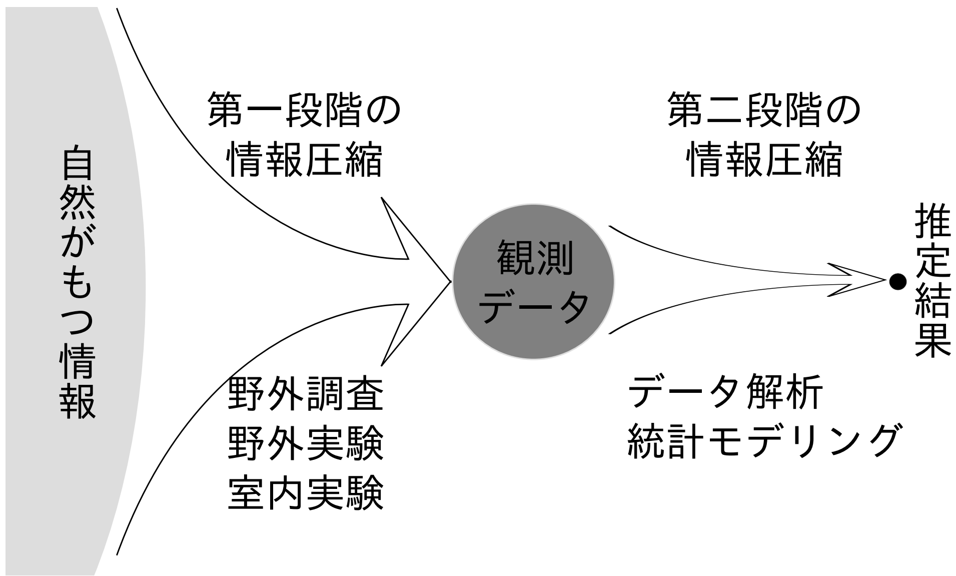
1. 実際のデータ解析で使う統計モデルの考えかたを
2. 基礎から最先端につながるような「地図」を示しつつ
3. 理解できる統計学めざして

なぜ私たちは統計学的な考えかたを勉強する必要があるのか?

自然科学研究における二段階の情報損失

第一段階: 自然現象 → 数値データ

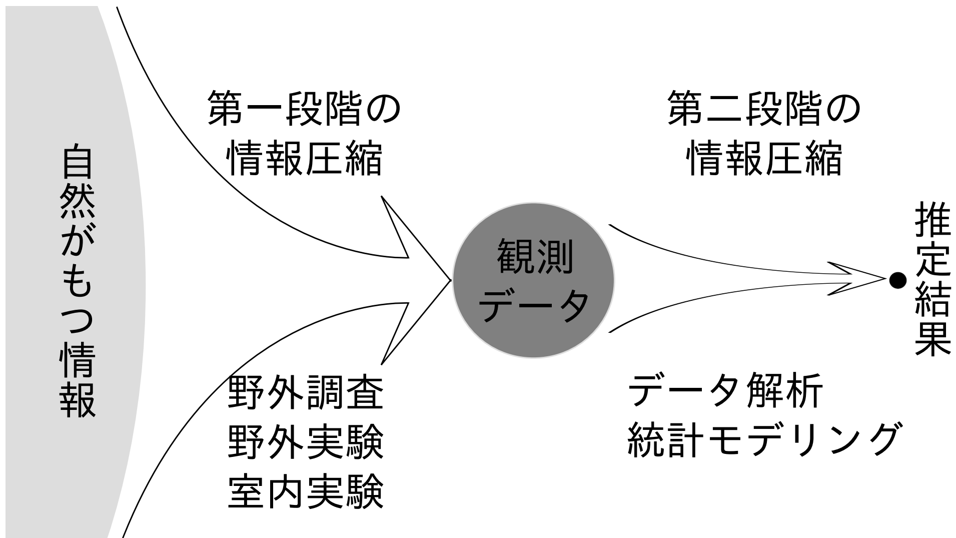
- 観察・実験による情報損失
- 人間が自然現象からとりだせる数値データはごくわずか
- (とくに野外調査では) 厳密に「同じ」データを再びとれない



自然科学研究における二段階の情報損失

第二段階: 数値データ → 統計学的な解析結果

- 統計解析による情報損失
- 人間のアタマは大量の数値データも把握できない
- この情報損失過程には再現性がある(“客観的” に検討できる)



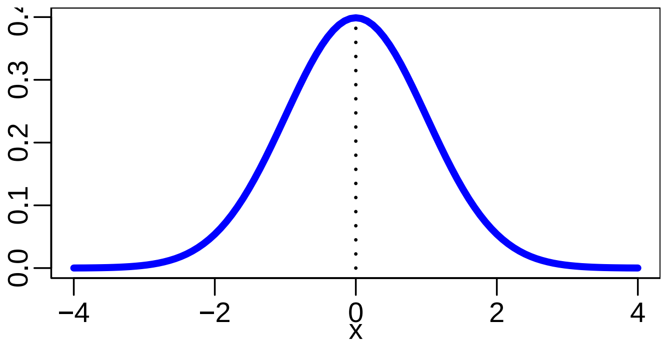
自然科学の研究をするためには，この過程のしくみを理解する必要がある

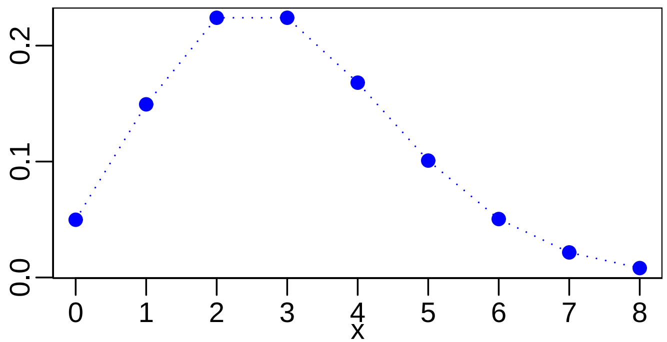
自然科学ではばらつきのある自然現象を……

背後にある確率論的モデルによって生成された，と仮定する

直接は見えない世界

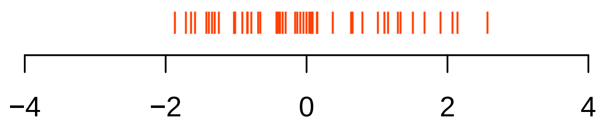
- モデル
- 確率分布
- 母集団

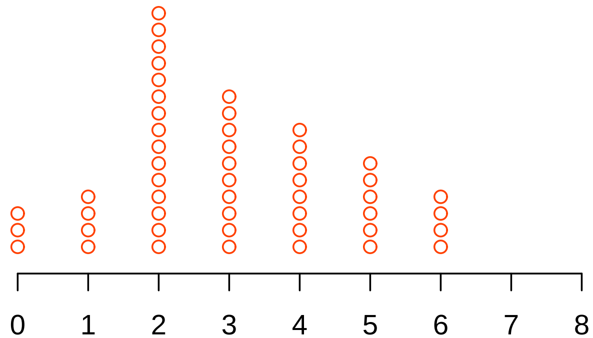




サンプリング $\Downarrow$ $\Uparrow$ (パラメーター) 推定

- データ
- 乱数
- 標本集団





見ることのできる世界

統計学的な解析の使われかたの現状

- 軽視されている (授業でも適切な方法を教えない)
- そもそも何やってるかわかってないヒトたちが多い
- まちがっている方法に固執する (指摘すると逆ぎれ)

理想 — この統計学授業のネラい

- 理念: スジのとおった合理的な統計解析 = 統計モデリングをめざそう
- 手段: データの性質・構造によくあった手法を (データの有効利用)
- 目的: 自然現象うまく説明できるモデリングになってれば

統計モデリングとはなんだろう?

統計モデリング: 観測データのモデル化

- 統計モデルは観測データのパターンをうまく説明できるようなモデル
- 基本的部品: 確率分布 (とそのパラメーター)
- データにもとづくパラメーター推定, あてはまりの良さを定量的に評価できる

「説明変数 $\rightsquigarrow$ 応答変数」型の統計モデルの地図 …… $\Longrightarrow$

ゆうど
尤度 をあつかう統計モデル

パラメーターを確率分布として表現する Bayes 統計学

階層 Bayes モデル の MCMC 計算による推定など

最尤推定法 であつかう統計モデル

パラメーターを点推定する，random effects もあつかえる

階層ベイズモデルである一般化線形混合モデル (GLMM) など

一般化線形モデル (GLM)

指数関数族の確率分布 + 線形モデル, fixed effects のみ

最小二乗法 であつかう統計モデル

等分散正規分布 + 線形モデル

直線回帰，いわゆる「分散分析」など

2. 統計ソフトウェア
R の簡単な紹介

教員世代より院生世代のほうが「わかっている」理由:
新しいデータ解析を学ぶときの勉強環境がすぐれている
(教員世代はもはや勉強しない …… ヒトが多い)

- 計算機のハードウェアがよい …… 格段に速い!
- 統計解析のソフトウェアがよい …… R!!
- インターネット上にいくらでも教材がある

皆さんは良い環境で勉強できる，ということです

これ使いましょう: 統計ソフトウェア R

http://www.r-project.org/

- いろいろな **OS** で使える **free software**
- 使いたい機能が充実している
- 作図機能も強力
- S 言語によるプログラミング可能
- よい教科書が出版されつつある
  - 「R による保健医療データ解析演習」 中澤港 (2007)
  - 「The R-Tips」 舟尾暢男 (2005)
  - "Statistics: An Introduction Using R" M. Crawley (2005)
  - ネット上のあちこち

R が変えつつある生態学のデータ解析

- 使いたい手法はたいていそろってる
- 無ければ自分で何でも簡単に作れる
- 統計学的 simulation も簡単にできる

…… となると ……

- データを無理やりある手法にこじつける，ということが不要になる— データの構造にあわせた統計モデリングを行えばよい
- 手法の前提となる統計学の基本(統計モデル) の理解がむしろ重要
- 単純な検定ではなく,「こういう標本のばらつきを生成したメカニズム」の推定のよしあしが問われる

説明したい統計モデリングのお作法

- 観測データの図をたくさん作ろう
- 観測データをどんな確率分布で表現できるか考えよう
- 「割算値」の統計モデリングはやめよう

つまり観測データの「もち味をいかした」
「ひねくりまわさない」統計モデリング

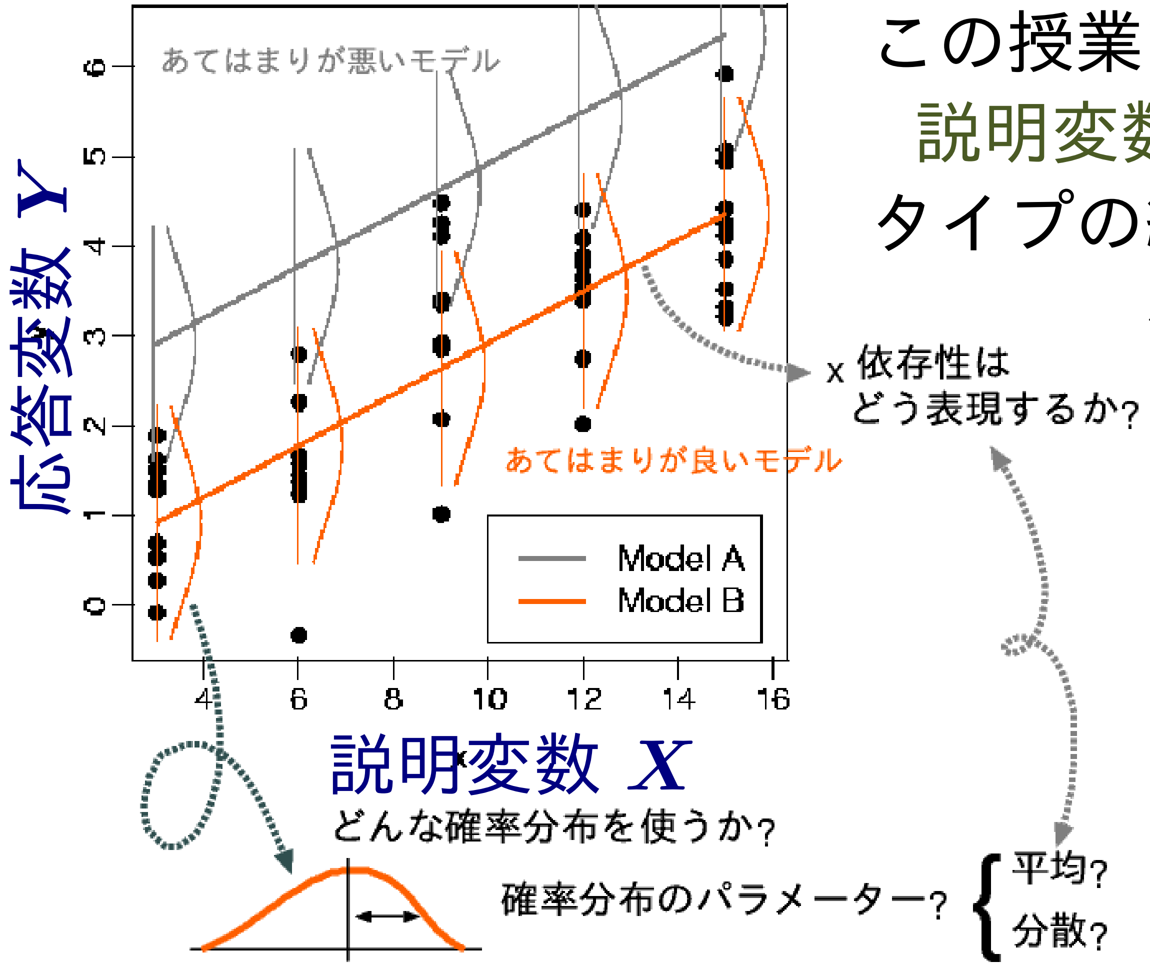
3. 一般化線形モデル: ポアソン回帰

統計モデリングとは何か?

データ解析とは統計モデリングのことだ

- 統計モデルは (解析したい) 観測データと対象に関する先験的な知識・情報にもとづいて構築される
- 統計モデルは観測データのパターンをうまく説明できるようなモデル
- 統計モデルの基本的な部品は確率分布，確率分布のカタチはパラメーターによって決まる
- 観測データをうまく説明できるようにパラメーターの値を決めることを「統計モデルのあてはめ」または「統計モデルによる推定」という

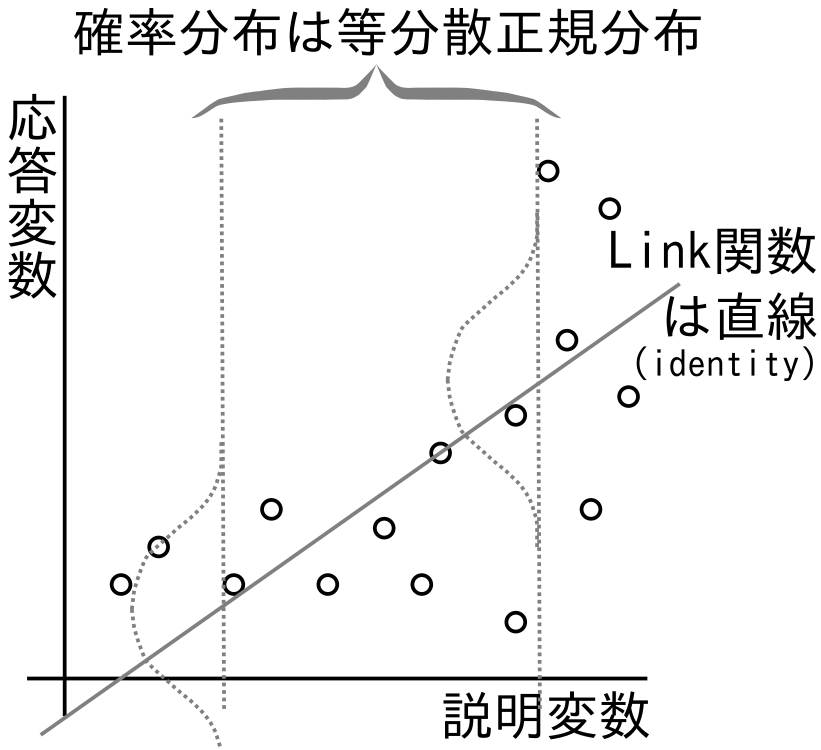
統計モデルの部品: 確率分布・パラメーター・…



この授業では
説明変数 ⇝ 応答変数
タイプの統計モデルのみ
をあつかう

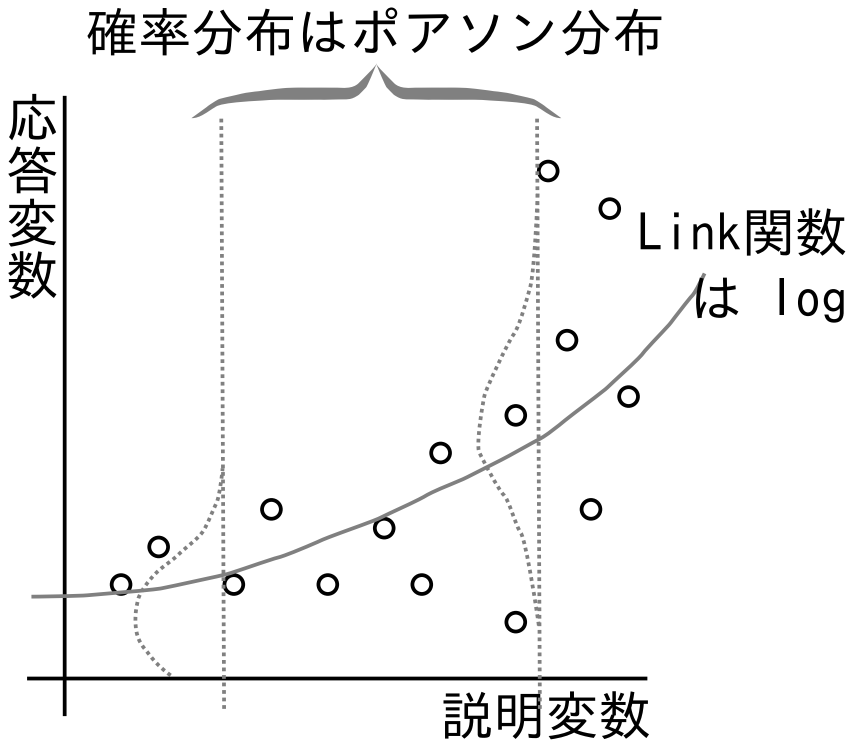
観測データと統計モデルの「比較」 ⇒ パラメーターの推定

統計モデル: いつでも「直線回帰」でいいのか?



- もしこの観測データ (縦軸) がカウントデータだったら?
- まずい点: 等分散ではないに直線回帰?
- まずい点: モデルによる予測は「負の個体密度」?

カウントデータならポアソン回帰で!



- ポアソン回帰は一般化線形モデルの一部
- 平均値とともに増大する分散に対応
- モデルによる予測はつねに非負

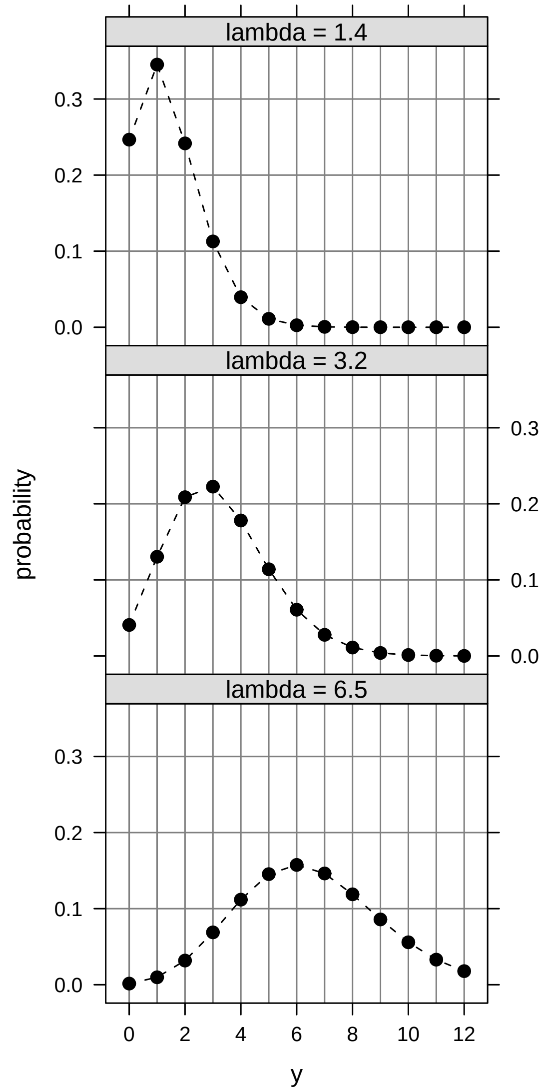
ポアソン分布 (Poisson distribution) とは何か?

- 離散分布 $y_i \in \{0, 1, 2, \cdots, \infty\}$
- 確率密度関数 (paramter: λ)

$$\frac{\lambda^y \exp(-\lambda)}{y!}$$

- 期待値 λ, 分散 λ

- 上限を設定できないカウントデータに
- 例: 産卵数・種子数・個体数……



ゆうど
尤度 をあつかう統計モデル

パラメーターを確率分布として表現する Bayes 統計学

階層 **Bayes** モデル の **MCMC** 計算による推定など

最尤推定法 であつかう統計モデル

パラメーターを点推定する，random effects もあつかえる

階層ベイズモデルである一般化線形混合モデル (GLMM) など

一般化線形モデル (GLM)

指数関数族の確率分布 + 線形モデル, fixed effects のみ

最小二乗法 であつかう統計モデル

等分散正規分布 + 線形モデル

直線回帰，いわゆる「分散分析」など

一般化線形モデル (generalized linear model; GLM)

確率分布・link 関数・線形予測子を
指定して特定できる統計モデル

- 確率分布: 応答変数のばらつきとして正規分布，ポアソン分布，二項分布その他を指定できる
- link 関数を $f()$ とすると，確率分布の平均値 $=$ f(線形予測子) という関係がある
- 線形予測子: $\beta_0 + \beta_1 x_1 + \beta_2 + \cdots$，ただし x_i は説明変数で β_i は x_i の係数 (coefficient)
  – 観測データ ($\{x_i\}$ と $\{y_i\}$) にもとづいて $\{\beta_i\}$ を最尤推定するのが，GLM によるパラメーター推定

R で一般化線形モデル: glm() 関数

	確率分布	乱数生成	パラメーター推定
(離散)	ベルヌーイ分布	`rbinom()`	`glm(family = binomial)`
	二項分布	`rbinom()`	`glm(family = binomial)`
	ポアソン分布	`rpois()`	`glm(family = poisson)`
	負の二項分布	`rnbinom()`	`glm.nb()`
(連続)	ガンマ分布	`rgamma()`	`glm(family = gamma)`
	正規分布	`rnorm()`	`glm(family = gaussian)`

- `glm()` で使える確率分布は上記以外もある
- `glm.nb()` は `MASS` **library** 中にある
- **GLM** は直線回帰・重回帰・分散分析・ポアソン回帰・ロジスティック回帰その他の「よせあつめ」と考えてもよいかも

R の glm() 関数: 何を指定すればいい?

結果を格納するオブジェクト

モデル式

関数名

確率分布の指定

```
fit <- glm(
  y ~ log.x,
  family = poission(link = "log")
  data = d
)
```

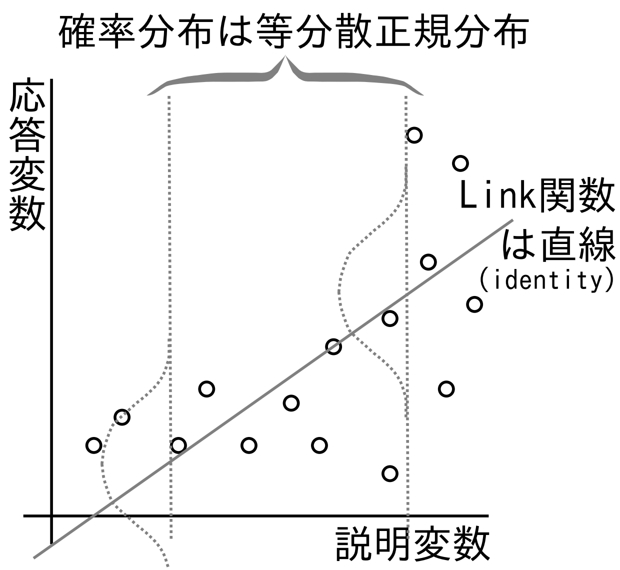
data.frame の指定

リンク関数の指定（省略可）

- モデル式 (線形予測子 z): どの説明変数を使うか?
- `link` 関数: z と応答変数 (y) 平均値 の関係は?
- `family`: どの確率分布を使うか?

「直線回帰」の `glm()` 指定 (1)

- `family`: `gaussian`, 正規分布
  - 本来は y が連続かつ $[-\infty, \infty]$
- `link` 関数: `"identity"`
  - これは `family = gaussian` 時の「おススめ」 **link** 関数

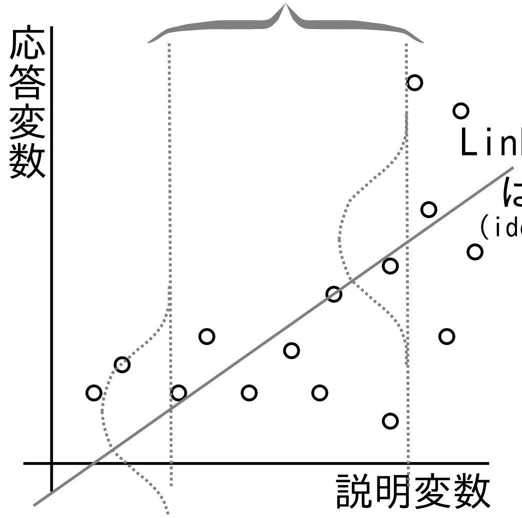


- モデル式 (線形予測子 z): たとえば `y ~ x` と指定したとする

データの点を見ても「正規分布」とは思えないのだけど，とりあえず `glm()` 指定の例を考えている

「直線回帰」の glm() 指定 (2)

- `family`: `gaussian`, 正規分布
- `link` 関数: `"identity"`
- モデル式 (線形予測子 z): たとえば `y ~ x` と指定したとする



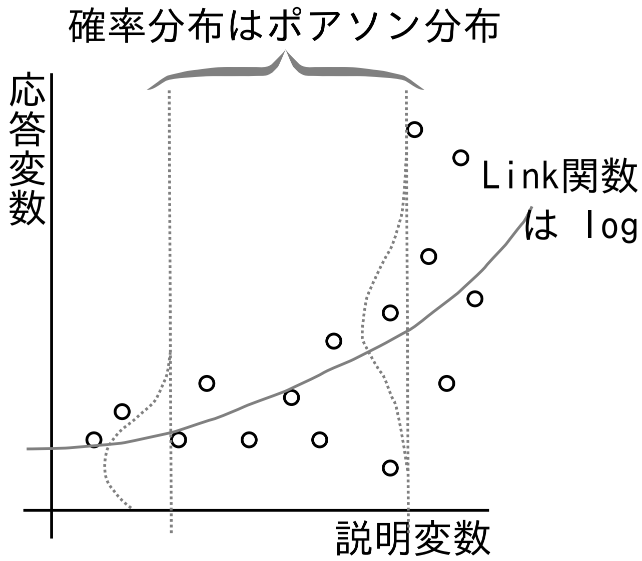
- 線形予測子 $z = a + bx$

  a, b は推定すべきパラメーター
- 応答変数の平均値を μ とすると $\mu = z$

  つまり $\mu = z = a + bx$
- 応答変数 は平均 μ の正規分布に従う: $y \sim \mathrm{Norm}(\mu, \sigma)$

ポアソン回帰の glm() 指定 (1)

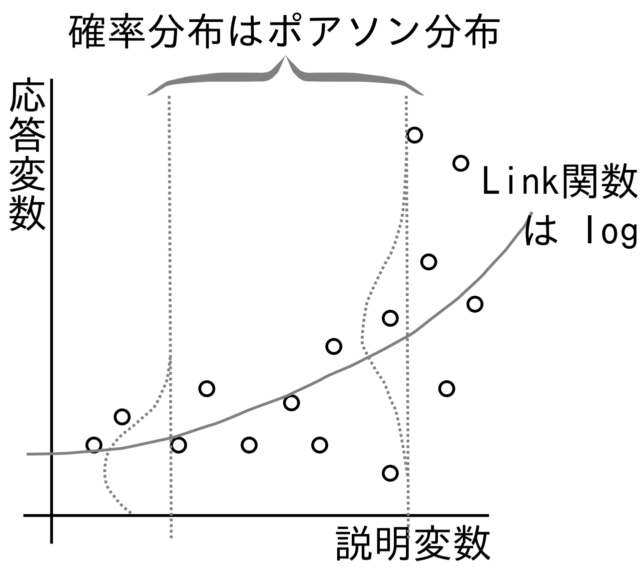
- family: poisson, ポアソン分布
  - カウントデータ $(0, 1, 2, \cdots$ と数えられるデータ) の場合はポアソン分布で説明してみる
- link 関数: "log"
  - これは family = poisson 時の「おススメ」 link 関数
- モデル式 (線形予測子 z): たとえば y ~ x と指定したとする



family = poisson(link = "log")
指定とは何をやっているのだろうか?

ポアソン回帰の `glm()` 指定 (2)

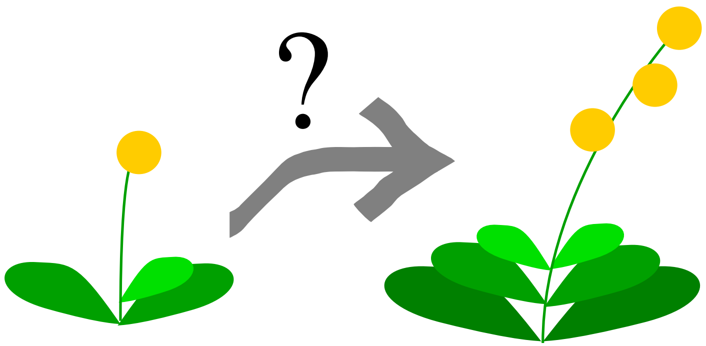
- `family`: `poisson`, ポアソン分布
- `link` 関数: `"log"`
- モデル式 (線形予測子 z): たとえば `y ~ x` と指定したとする

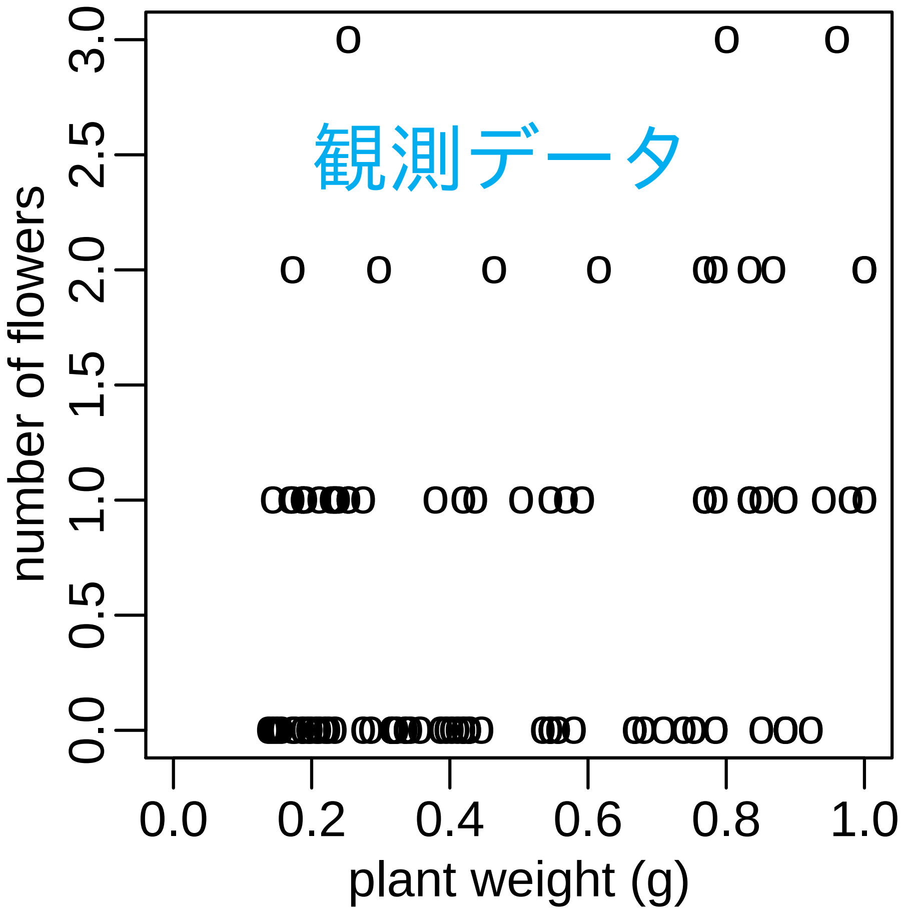


- 線形予測子 $z = a + bx$
  a, b は推定すべきパラメーター
- 応答変数の平均値を λ とすると $\log(\lambda) = z$
  つまり $\lambda = \exp(z) = \exp(a + bx)$
- 応答変数 は平均 λ のポアソン分布に従う: $y \sim \text{Pois}(\lambda)$

今日の例題: サイズと花数の関係?

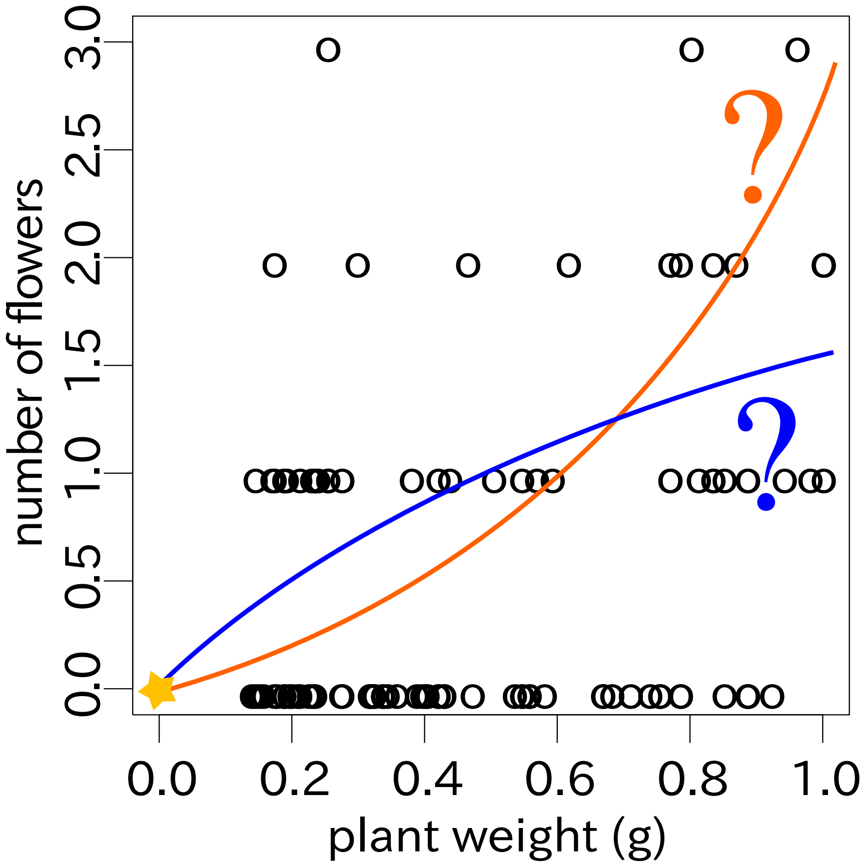
地上部の重量 x
が増加するにつれて
花数 y は増加する
だろうか?





- 調べた個体数は 100 個体: $i = 1, 2, \cdots, 100$
- 説明変数は地上部の重量 x_i
- 応答変数は花数 y_i

統計モデリング: x と y の関係は?



- 原点 $(0, 0)$ はとおる，と仮定しよう……
- 関数型は急上昇? 比例? アタマうち?

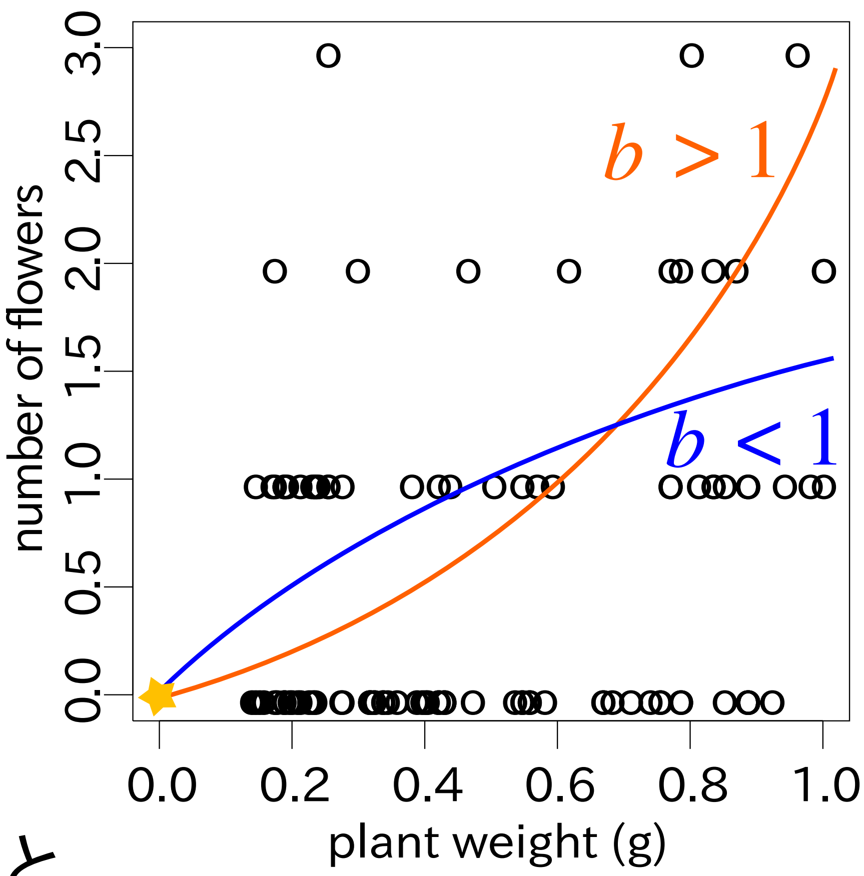
“アロメトリック” なモデルが良さそう

1. 応答変数 y_i は平均 λ_i のポアソン分布にしたがうと仮定:

$$y_i \sim \text{Pois}(\lambda_i)$$

2. ポアソン分布の平均 λ_i は x_i のベキ関数であると仮定:

$$\lambda_i = A x_i^b$$



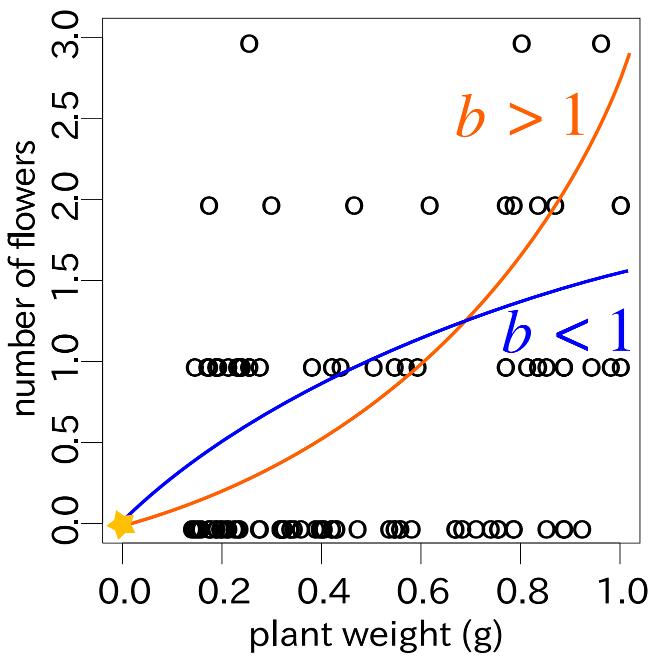
$\lambda_i = A x_i^b$ を変形してみると

$$\lambda_i = \exp(\log(A) + b \times \log(x_i))$$

$a = \log(A)$ とすると，$\log(\lambda_i) = a + b \times \log(x_i)$

この問題は GLM であつかえる!

- `family`: `poisson`, ポアソン分布
- `link` 関数: `"log"`
- モデル式 : `y ~ log.x` と指定，ただし重量 `x` の対数を `log.x` する



- 線形予測子 $z = a + b$ `log.x`

  a, b は推定すべきパラメーター
- 応答変数の平均値を λ とすると $\log(\lambda) = z$

  つまり $\lambda = \exp(z) = \exp(a + b$ `log.x`$)$
- 応答変数 は平均 λ のポアソン分布に従う: $y \sim \text{Pois}(\lambda)$

R に格納されたデータセットを操作する

編集前の data.frame “d” ⟹ `log.x` 列を追加する

```
> load("d.RData")
> head(d) # 先頭 6 行の表示
        x y
1 0.66762 0
2 0.85077 0
3 0.68124 0
4 0.14379 1
5 0.25316 1
6 0.88585 0
```

```
> d$log.x <- log(d$x)
> head(d)
        x y    log.x
1 0.66762 0 -0.40404
2 0.85077 0 -0.16162
3 0.68124 0 -0.38384
4 0.14379 1 -1.93939
5 0.25316 1 -1.37374
6 0.88585 0 -0.12121
```

glm() 関数の指定

結果を格納するオブジェクト

関数名

モデル式

確率分布の指定

```
fit <- glm(
  y ~ log.x,
  family = poission(link = "log")
  data = d
)
```

リンク関数の指定（省略可）

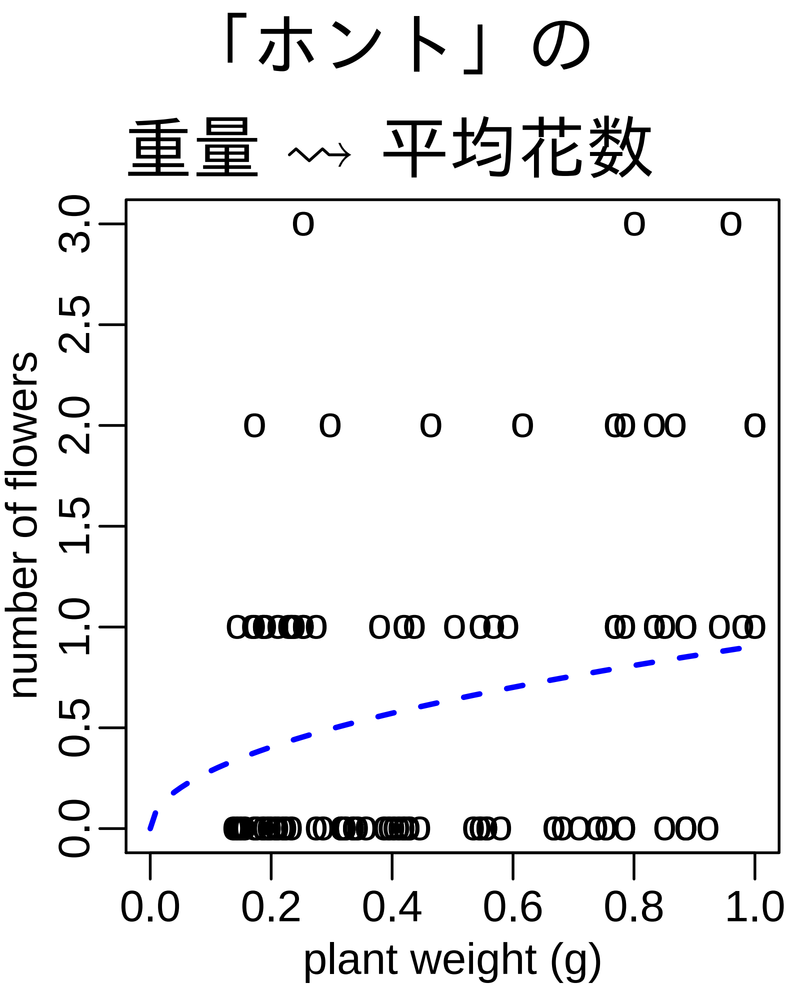
data.frame の指定

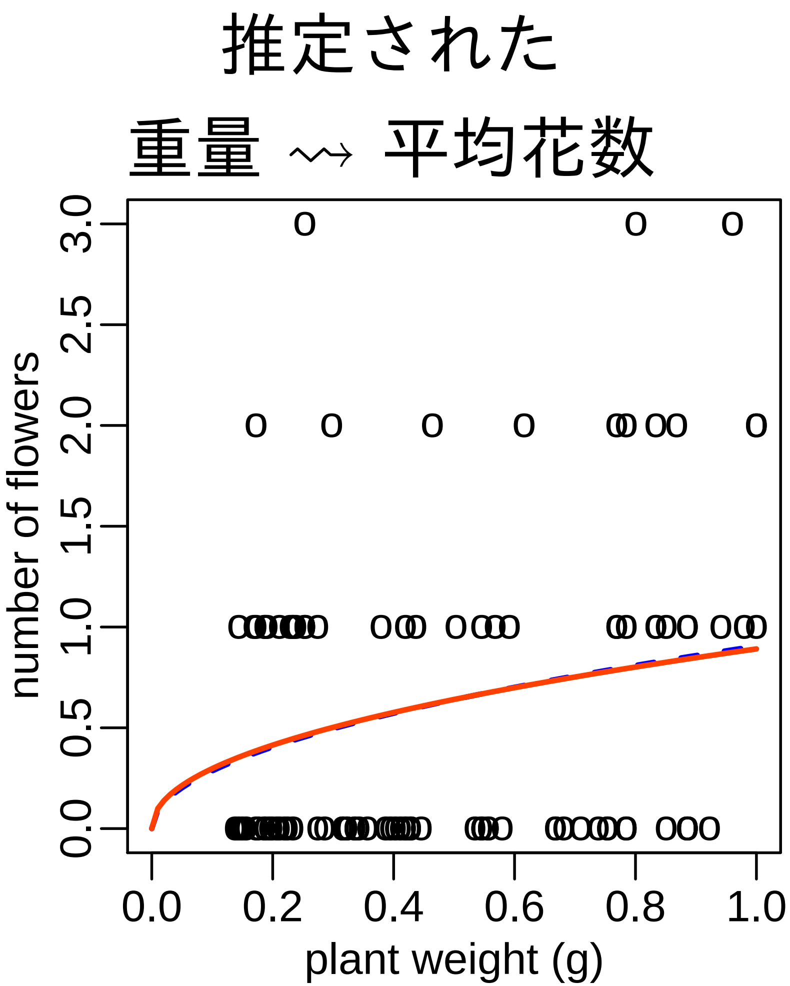
R の glm() 関数による推定結果

```
> fit <- glm(y ~ log.x, data = d, family = poisson)
> print(summary(fit))

Call:
glm(formula = y ~ log.x, family = poisson, data = d)
(... 略...)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.115      0.204   -0.56    0.573
log.x          0.476      0.222    2.14    0.032
(... 略...)
```

Coefficients は説明変数の係数という意味

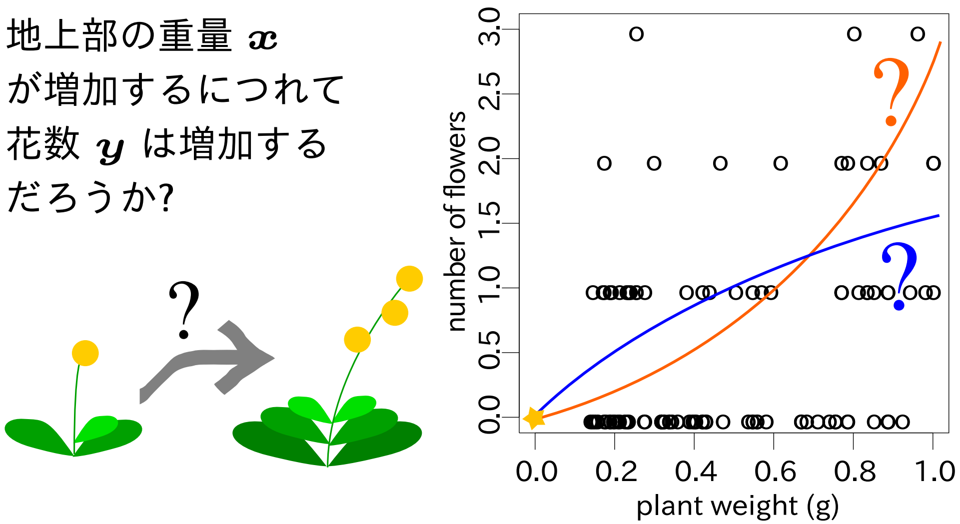
GLM の推定結果を図示してみる





今日のハナシ

1. データ解析は統計モデリングだ
2. 統計ソフトウェア R を使おう，作図重要
3. データをよくみて統計モデルの確率分布を選び，R の glm() を使いこなそう



説明したい統計モデリングのお作法

- 観測データの図をたくさん作ろう
- 観測データをどんな確率分布で表現できるか考えよう
- 「割算値」の統計モデリングはやめよう

つまり観測データの「もち味をいかした」
「ひねくりまわさない」統計モデリング

次回予告

生態学基礎論 (生物多様性論 II) 2008-01-28

全部で 2 回講義の 2

一般化線形モデル (GLM) の基礎

なんでも「割算」するな!

「観測値わる観測値」な統計解析をやめる方法

http://hosho.ees.hokudai.ac.jp/~kubo/stat/