

データ解析で出会う統計的問題

GLM から始める統計モデリング

1. **GLM** の部品: 確率分布, link 関数, 線形予測子, ~~deviance~~

久保拓弥 (20 分)

2. **GLM** の使用注意: 交互作用の意味

粕谷英一 (60 分)

3. **GLM** の部品 (続): 「脱」 割算値のための offset 項わざ

久保拓弥 (40 分)

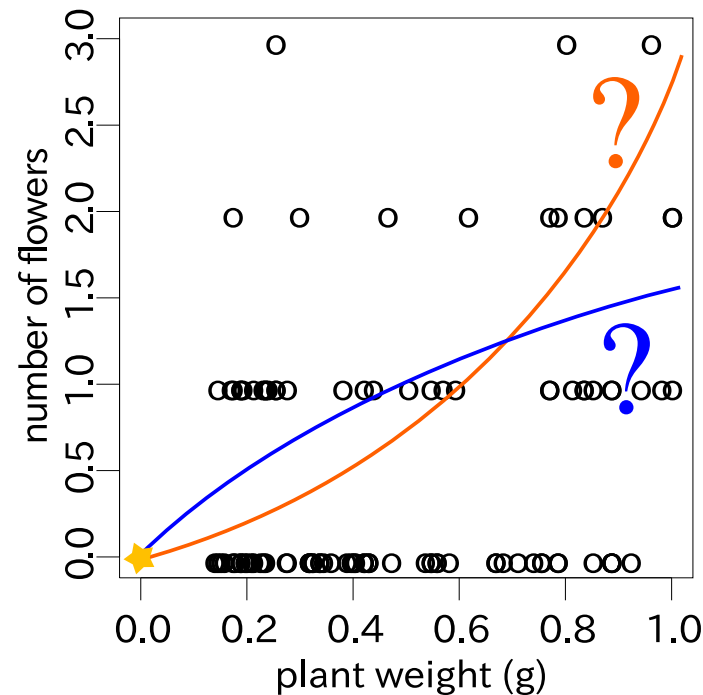
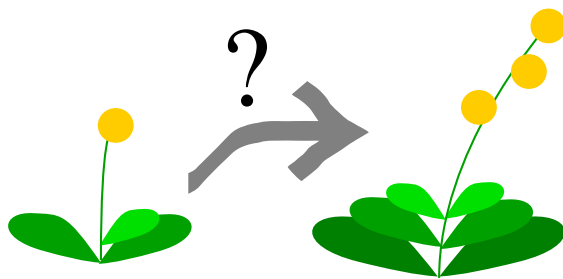
<http://hosho.ees.hokudai.ac.jp/~kubo/ce/EcoSj2009.html>

投影資料などダウンロードできます

今日のハナシ: 久保前半の内容

1. 統計モデルや GLM って何なの?
2. 今日の例題 A: ベキ関数 (power function) とポアソン回帰

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?



今日も統計ソフトウェア R 利用前提のハナシで

<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



1. 統計モデルや GLM って何なの？

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

今回は,「結果 ← 原因」関係が一番単純に表現している**線形モデル**のみを検討する

「結果 ← 原因」 関係を表現する線形モデル

- 結果: 応答変数
- 原因: 説明変数
- 線形予測子 (linear predictor):

$$\begin{aligned} \text{(応答変数の平均)} &= \text{定数 (切片)} \\ &+ \text{(係数 1)} \times \text{(説明変数 1)} \\ &+ \text{(係数 2)} \times \text{(説明変数 2)} \\ &+ \text{(係数 3)} \times \text{(説明変数 3)} \\ &+ \dots \end{aligned}$$

(交互作用項については粕谷さんが説明してくれます)

線形モデルの拡張

階層ベイズモデル

事前分布・事後分布を明示的にあつかう

一般化線形混合モデル (GLMM)

GLM + random effects (個体差・場所差など)

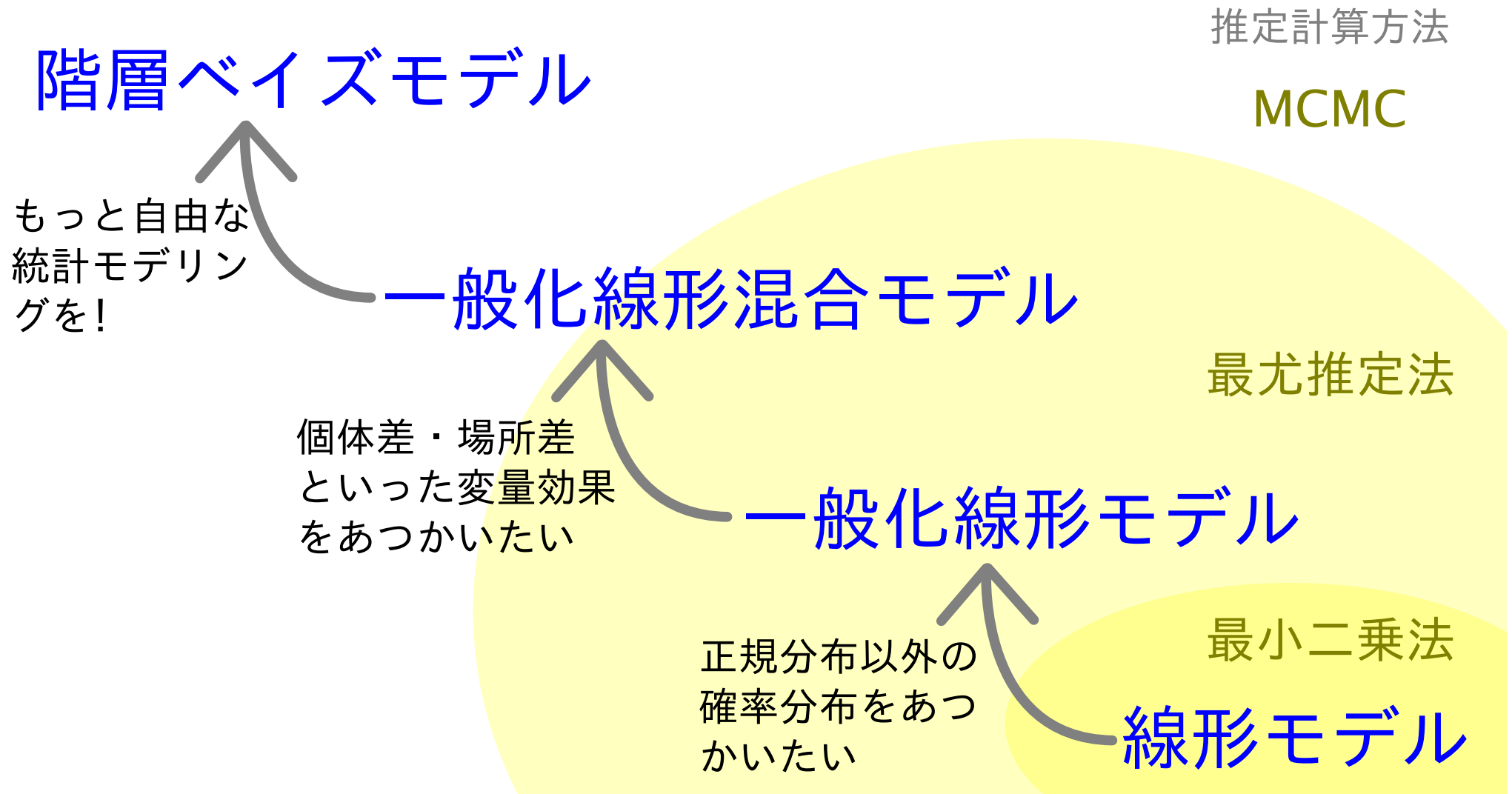
一般化線形モデル (GLM)

ポアソン回帰・ロジスティック回帰など

等分散正規分布 + 線形モデル

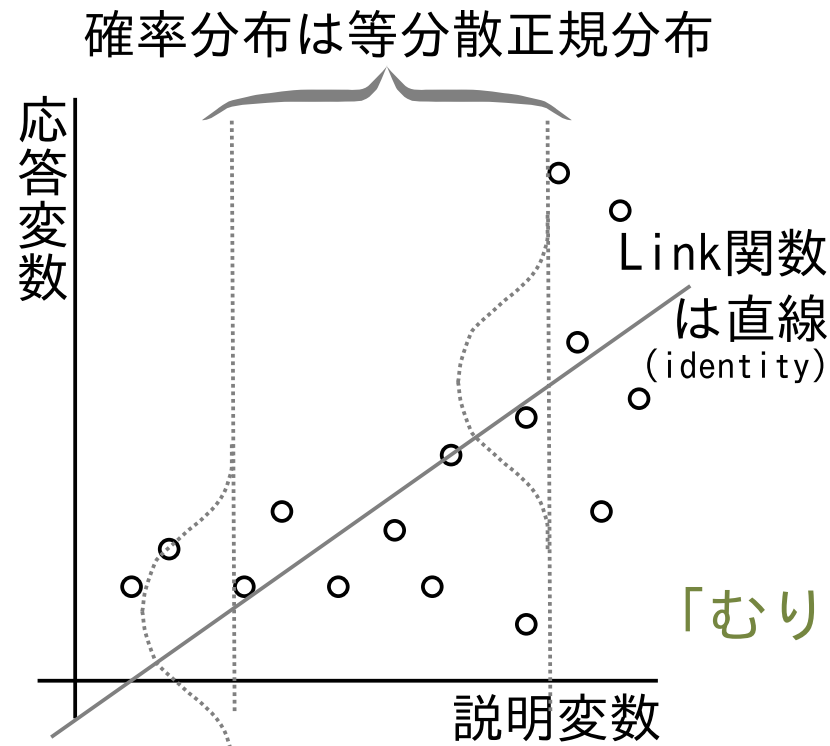
直線回帰, いわゆる「分散分析」など

線形モデルの発展



今日は「線形モデル → GLM」の発展を考えます

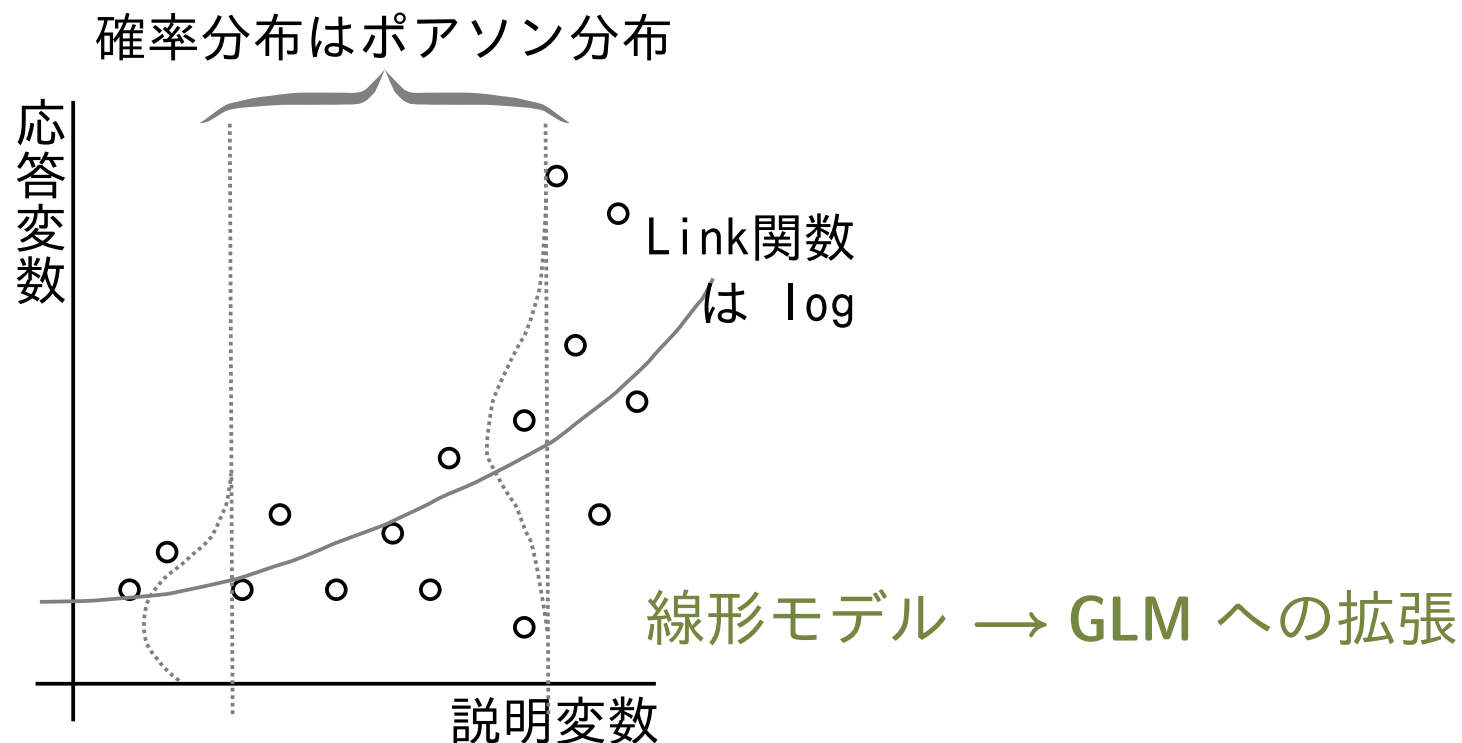
統計モデル: いつでも「直線回帰」でいいのか?



「むりやり線形モデル」の限界!

- 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!



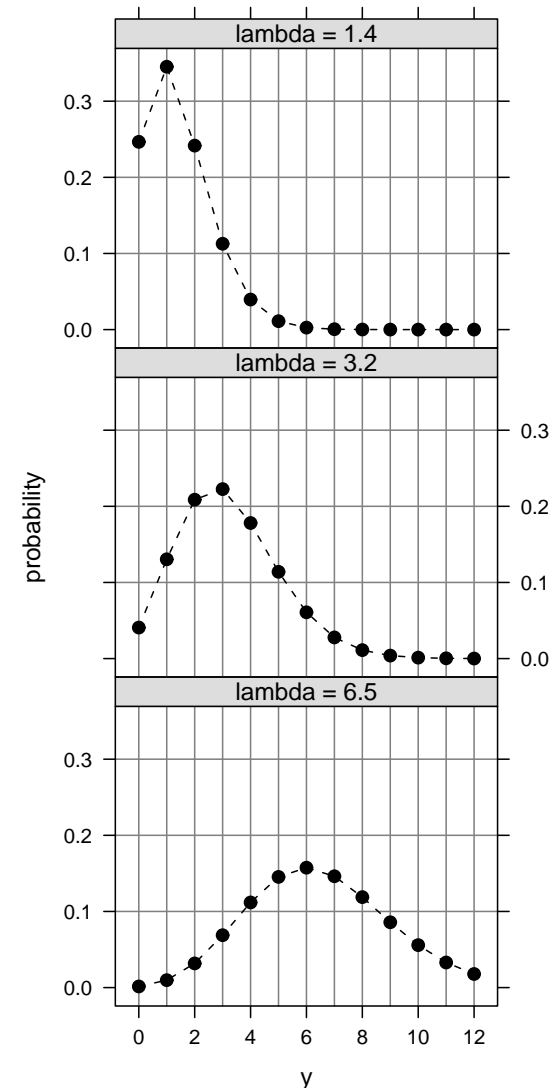
- ポアソン回帰は一般化線形モデルの一部
- 平均値とともに増大する分散に対応
- モデルによる予測はつねに非負

ポアソン分布 (Poisson distribution) とは何か?

- 離散分布 $y_i \in \{0, 1, 2, \dots, \infty\}$
- 確率密度関数 (parameter: 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……



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

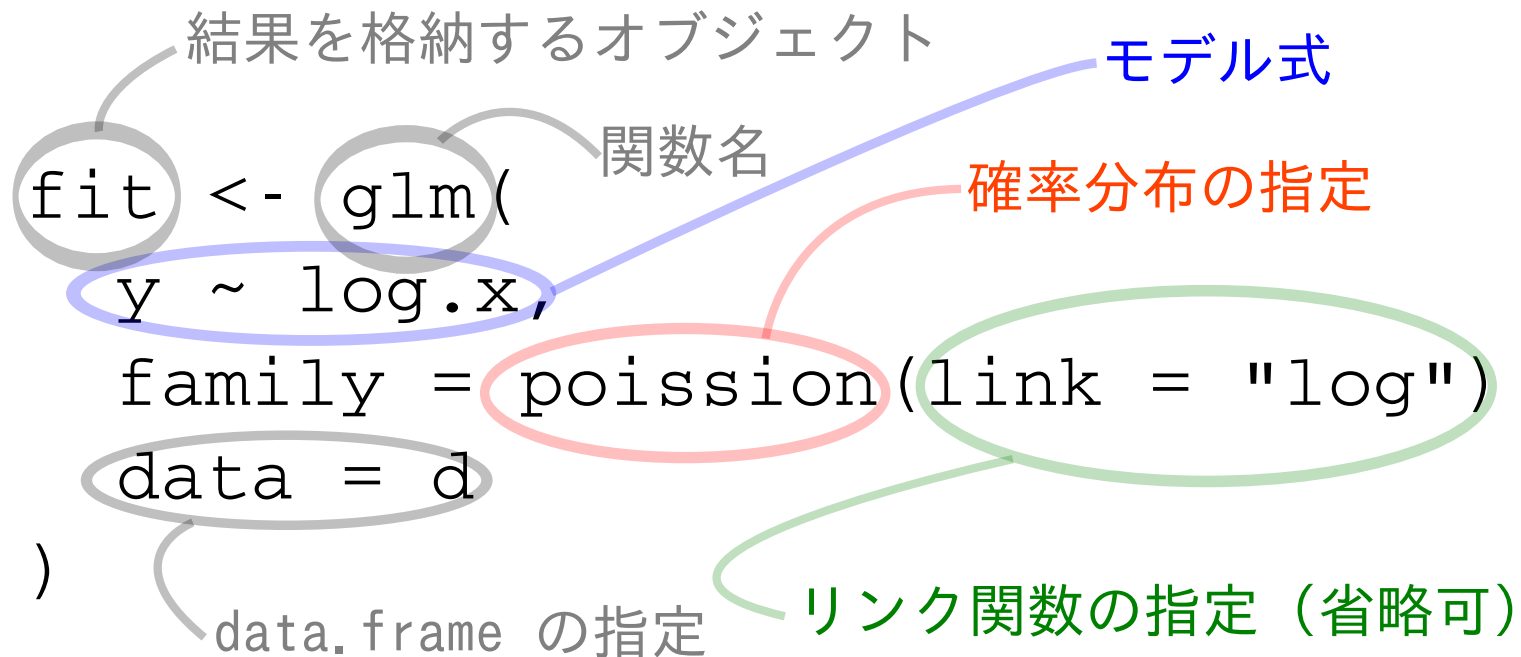
- 確率分布: 応答変数のばらつきとして正規分布, ポアソン分布, 二項分布その他を指定できる
- link 関数を $f()$ とすると, 確率分布の平均値 = $f(\text{線形予測子})$ という関係がある
- 線形予測子: $\beta_0 + \beta_1 x_1 + \beta_2 + \dots$, ただし x_i は説明変数で β_i は x_i の係数 (coefficient)
 - 観測データ ($\{x_i\}$ と $\{y_i\}$) にもとづいて $\{\beta_i\}$ を最尤推定するのが, GLM によるパラメータ推定

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() in library(MASS)
(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 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
- 今日はポアソン回帰を使った GLM だけ紹介します

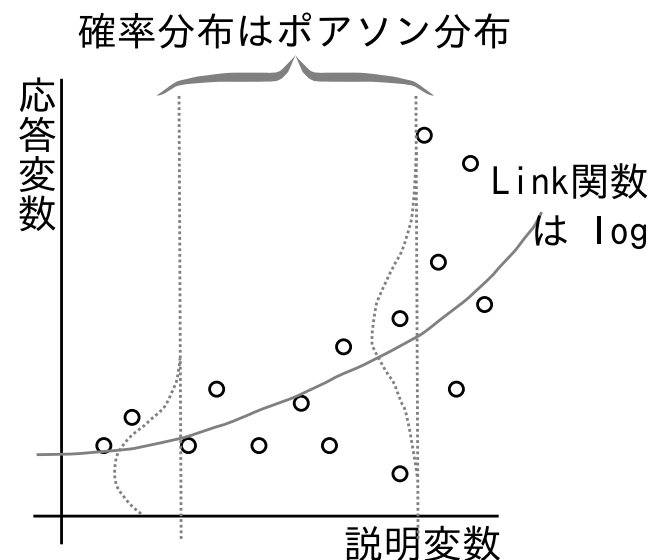
R の glm() 関数: 何を指定すればいい?



- 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?

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, ... 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 $y \sim x$ 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 $y \sim x$ と指定したとする

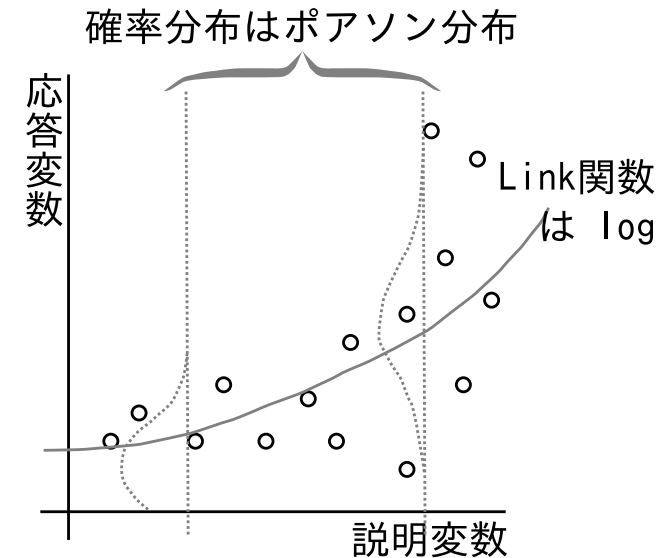
- 線形予測子 $z = a + bx$

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: $y \sim \text{Pois}(\lambda)$

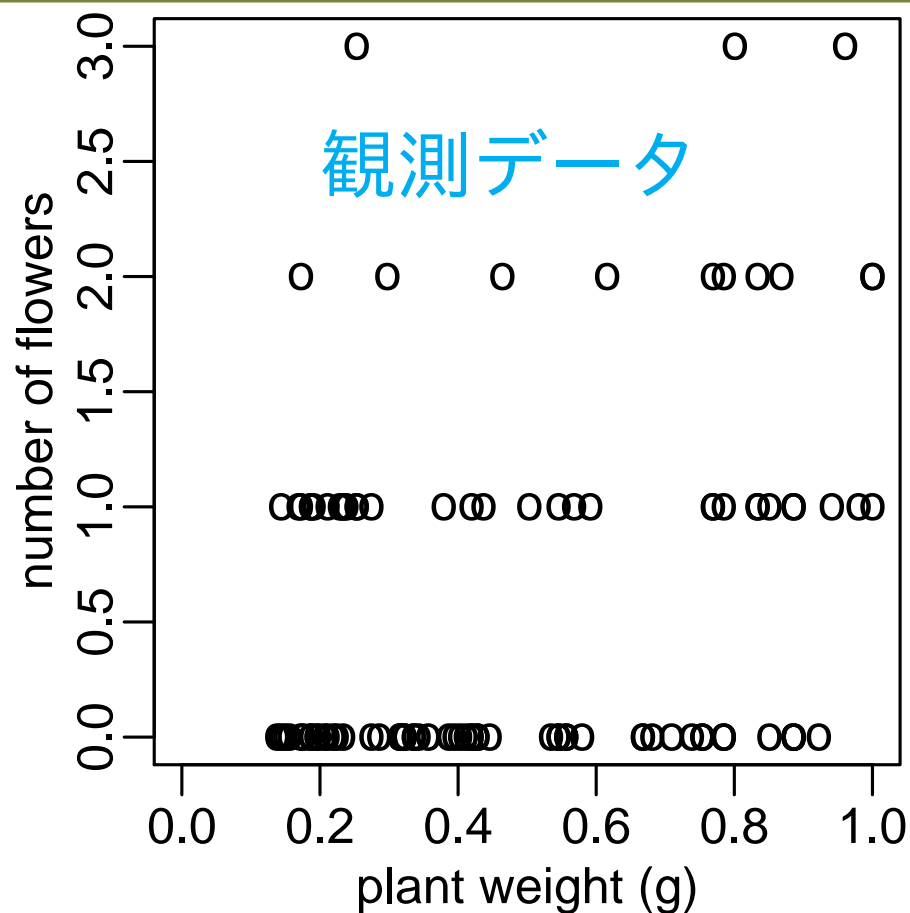
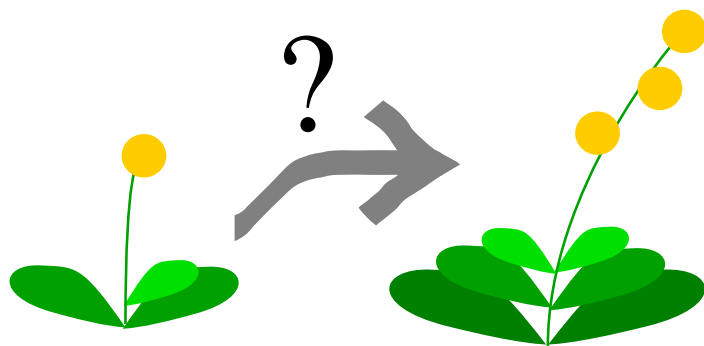


2. 今日の例題 A:

べき関数 (power function) とポアソン回帰

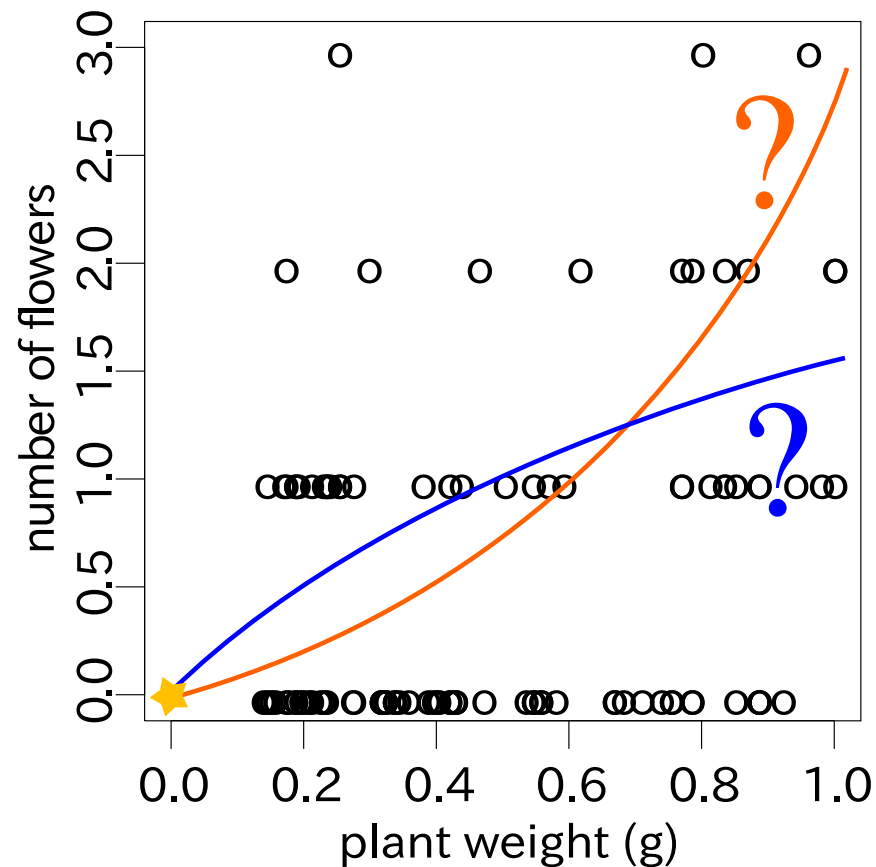
今日の例題 A: サイズと花数の関係?

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: $i = 1, 2, \dots, 100$
- 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?



- とりあえず「サイズ x とともに増大」と仮定し ……
- さらに原点 $(0, 0)$ はとおる, と仮定しよう ……
- とくに知りたいこと: 関数型は急上昇? 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:

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

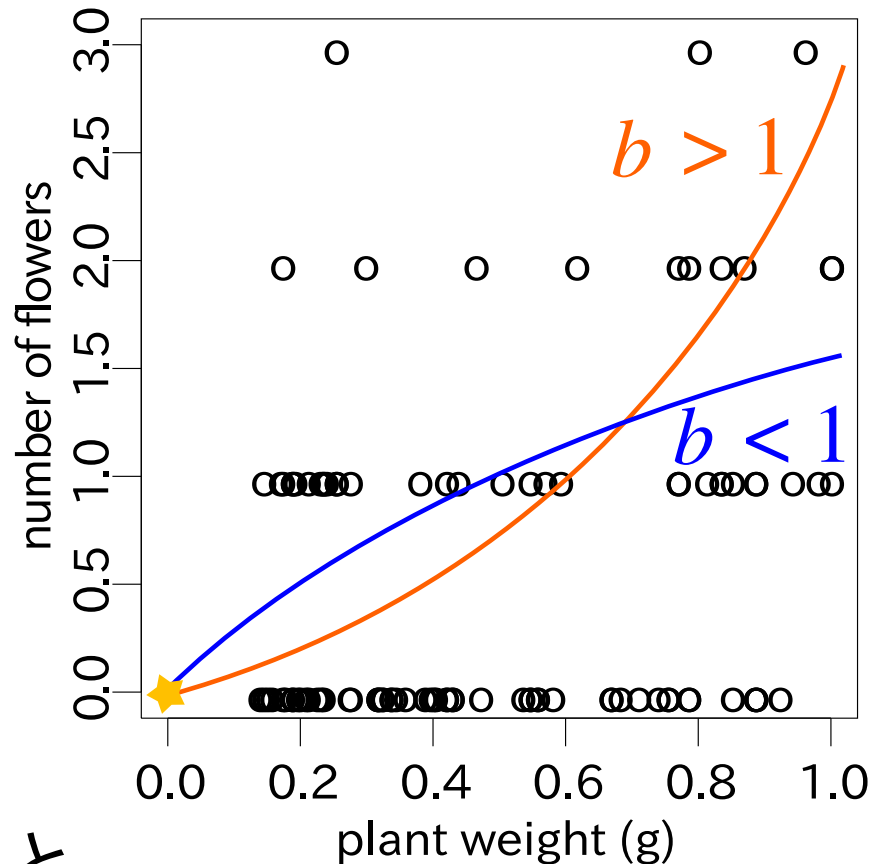
2. ポアソン分布の平均 λ_i は x_i のべき関数であると仮定:

$$\lambda_i = Ax_i^b$$

$\lambda_i = Ax_i^b$ を変形してみると

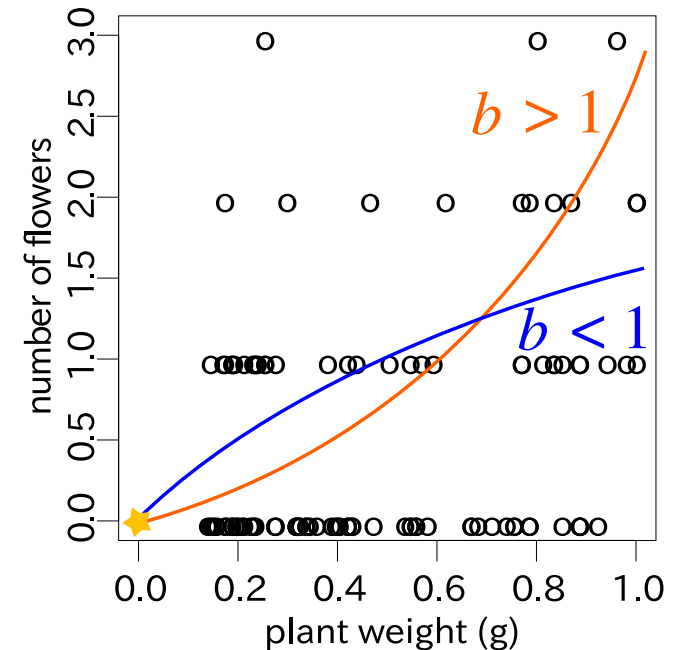
$$\lambda_i = \exp(\log(A) + b \times \log(x_i))$$

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



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: $y \sim \log.x$ 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” \implies log.x 列を追加する

```
> load("d.RData")
```

```
> head(d) # 先頭 6 行の表示
```

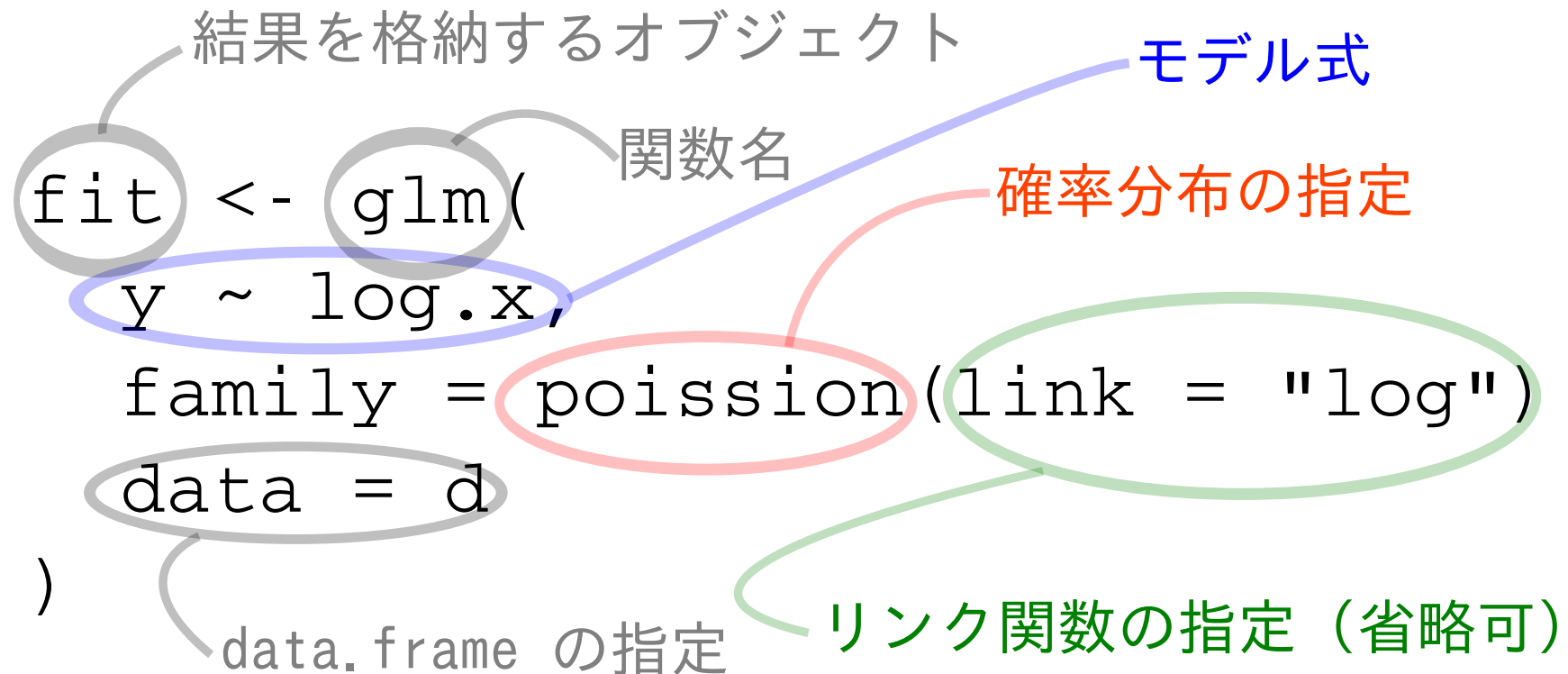
	x	y
1	0.66762	0
2	0.85077	0
3	0.68124	0
4	0.14379	1
5	0.25316	1
6	0.88585	0

```
> 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



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:

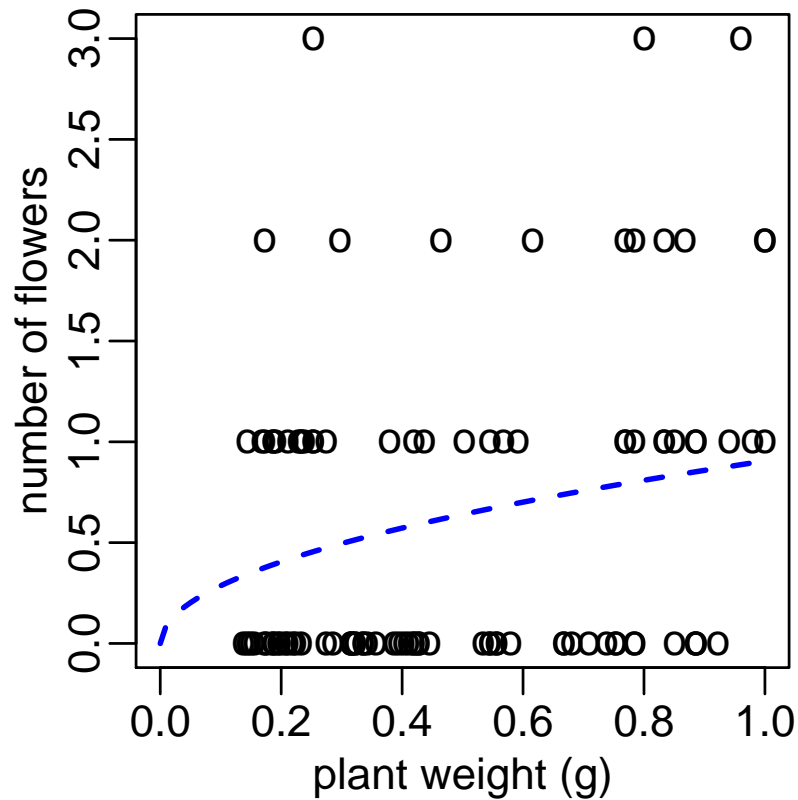
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.115	0.204	-0.56	0.573
log.x	0.476	0.222	2.14	0.032

(... 略...)

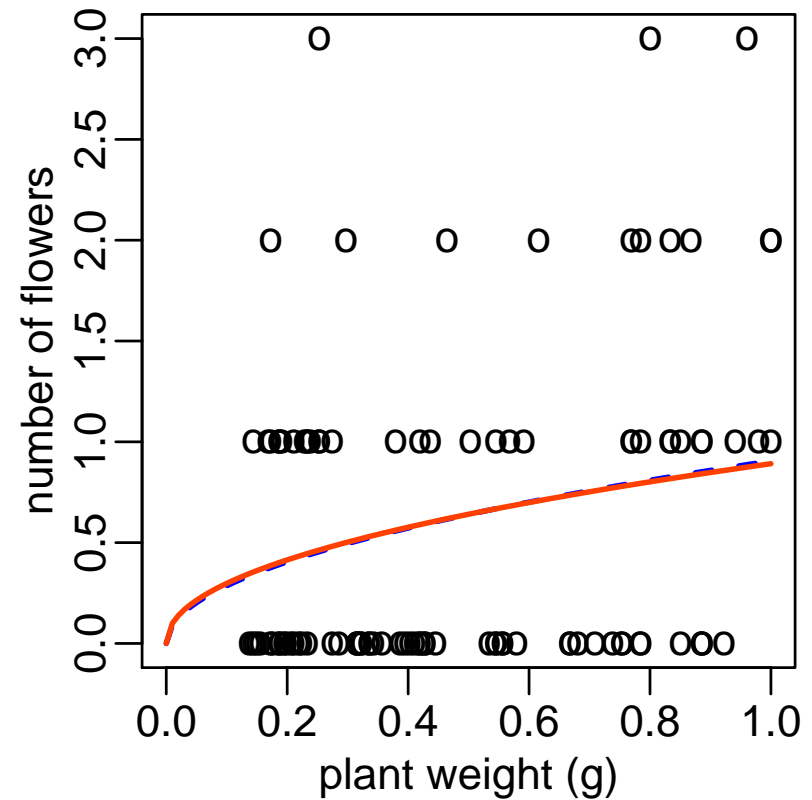
Coefficients は説明変数の係数という意味

GLM の推定結果を図示してみる

「ホント」の
重量 \rightsquigarrow 平均花数



推定された
重量 \rightsquigarrow 平均花数



今日のハナシ: 久保前半のまとめ

1. 統計モデル: 確率分布を使って観測データに見られるパターンを説明
2. GLM の部品: 確率分布, link 関数, 線形予測子
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

すみませんが, deviance や尤度のハナシはまた別の機会に……

久保前半のハナシはここまで

1. **GLM** の部品: 確率分布, link 関数, 線形予測子, ~~deviance~~

久保拓弥 (20 分)

2. **GLM** の使用注意: 交互作用の意味

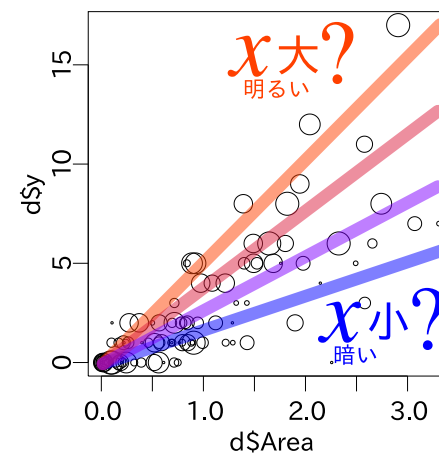
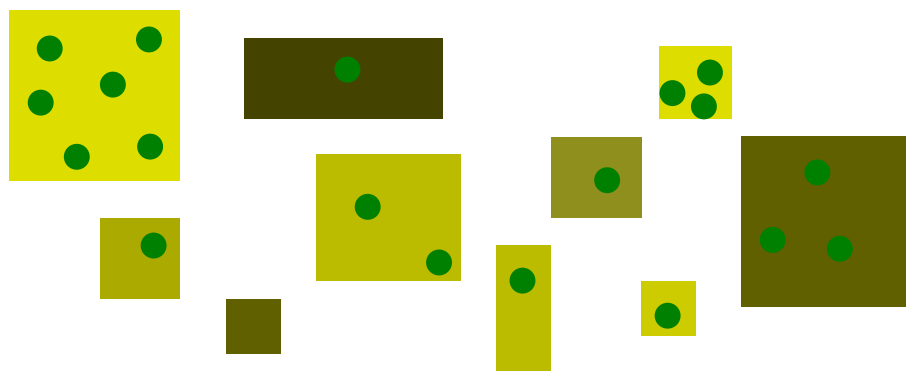
粕谷英一 (60 分)

3. **GLM** の部品 (続): 「脱」 割算値のための offset 項わざ

久保拓弥 (40 分)

今日のハナシ: 久保後半の内容

1. 割算解析やめましょう: その前に前半の復習
2. 「脱」割算の offset 項わざわざ: ポアソン回帰を強めてみる



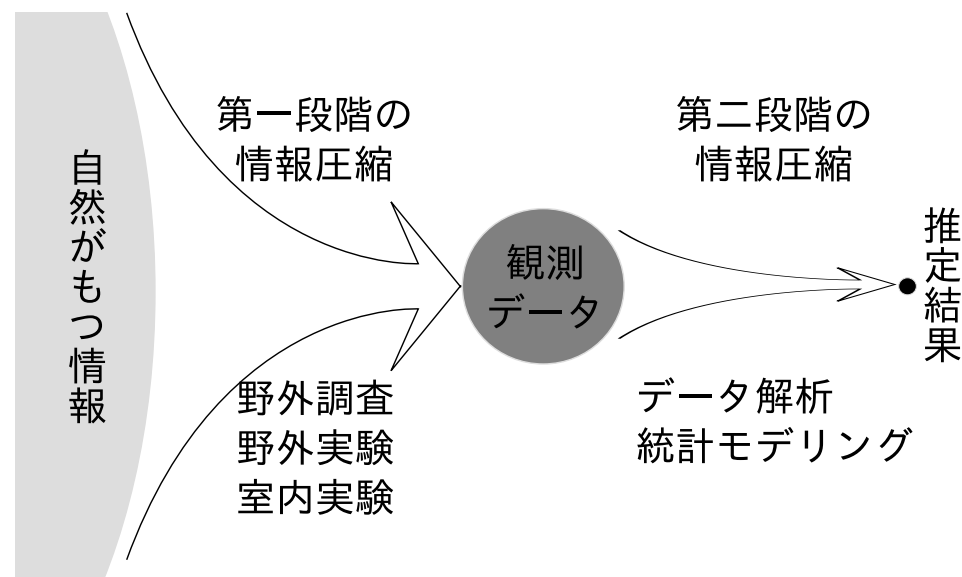
1. 割算解析やめましょう

その前に前半の復習

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

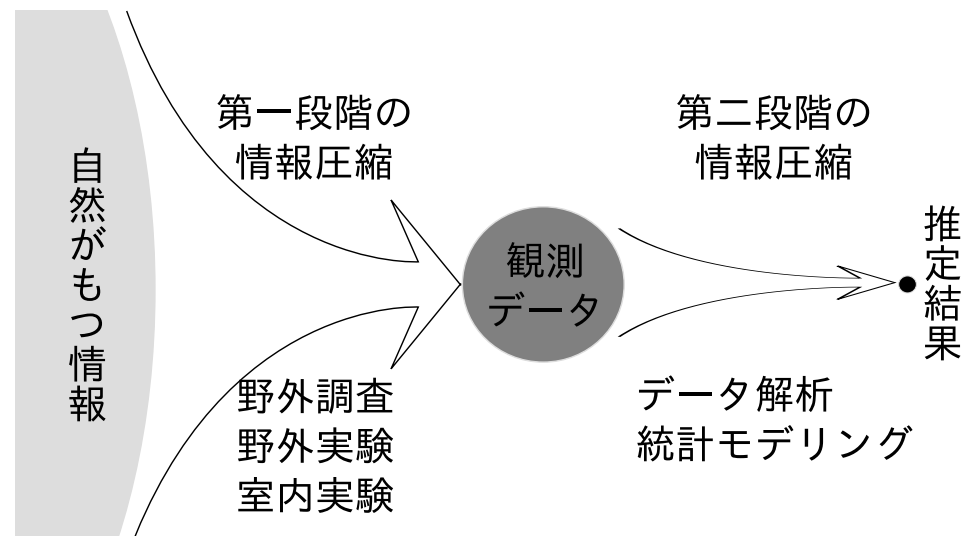
- 観察・実験による情報損失
- 人間が自然現象からとりだせる数値データはごくわずか
- (とくに野外調査では) 厳密に「同じ」データを再びとれない



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)



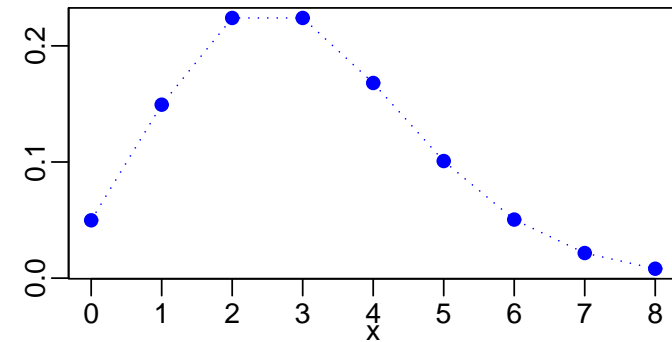
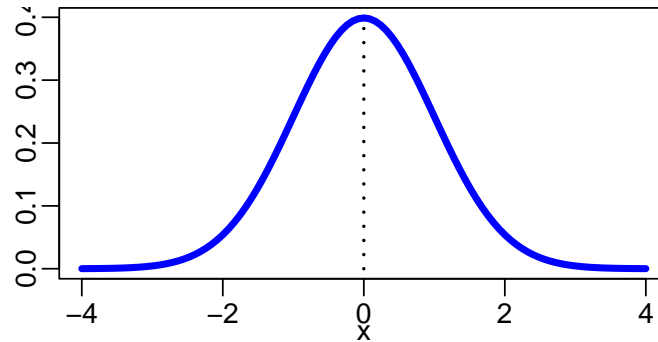
自然科学の研究をするためには、この過程のしくみを理解する必要がある

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

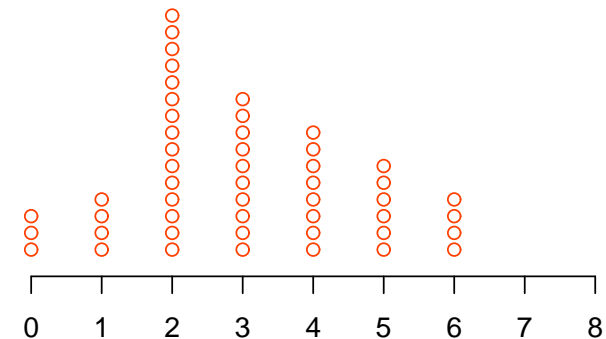
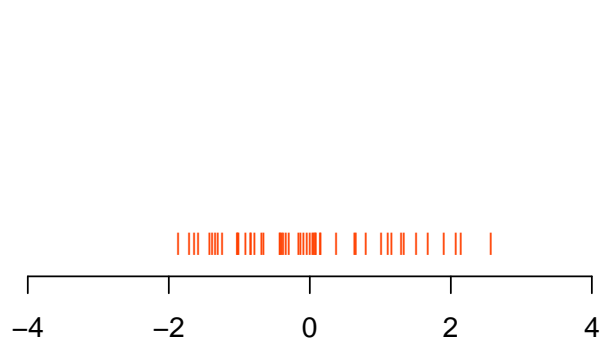
直接は見えない世界

- モデル
- 確率分布
- 母集団



サンプリング ↓ ↑ (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

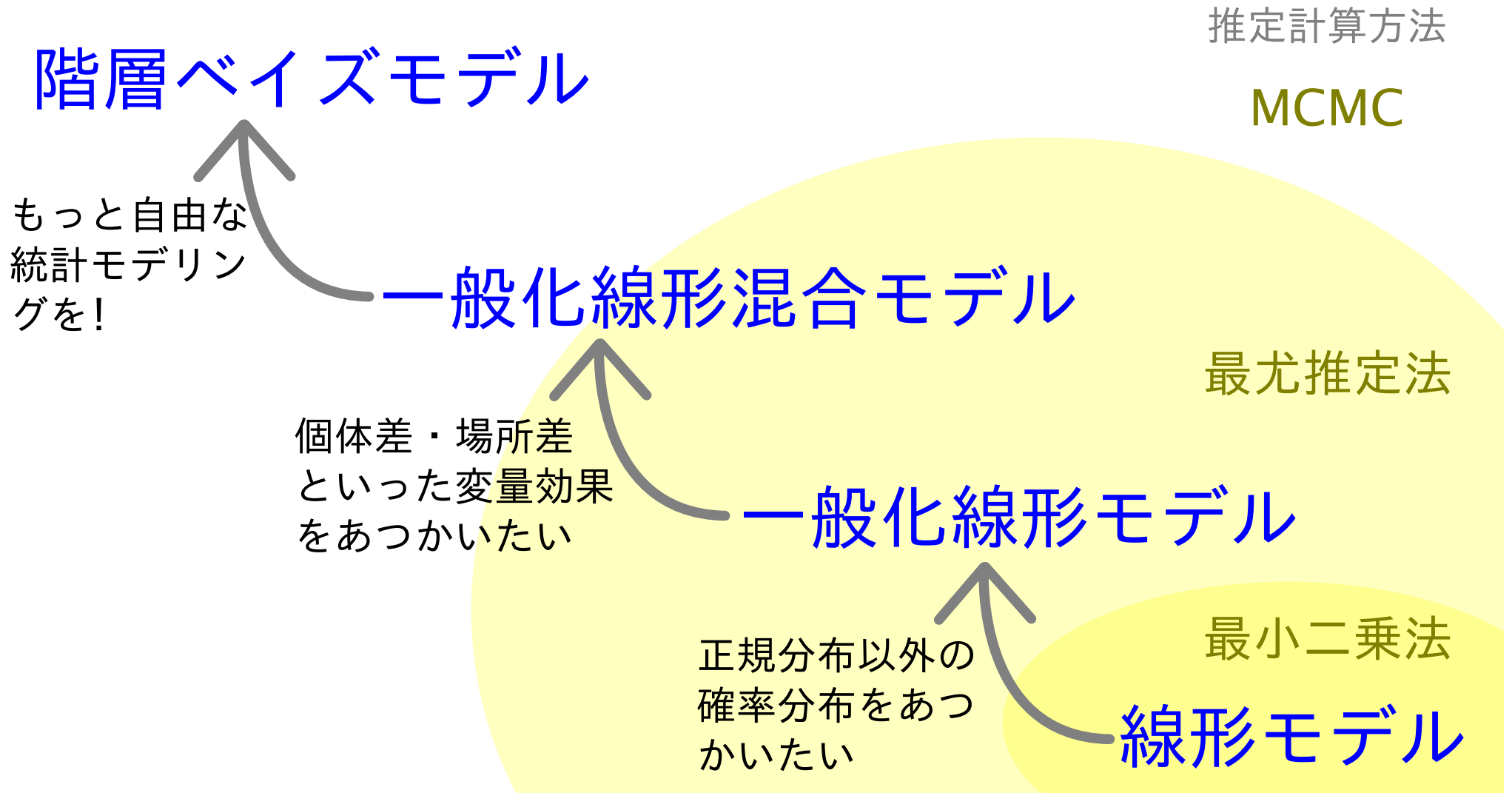
「結果 ← 原因」 関係を表現する線形モデル

- 結果: 応答変数
- 原因: 説明変数
- 線形予測子 (linear predictor):

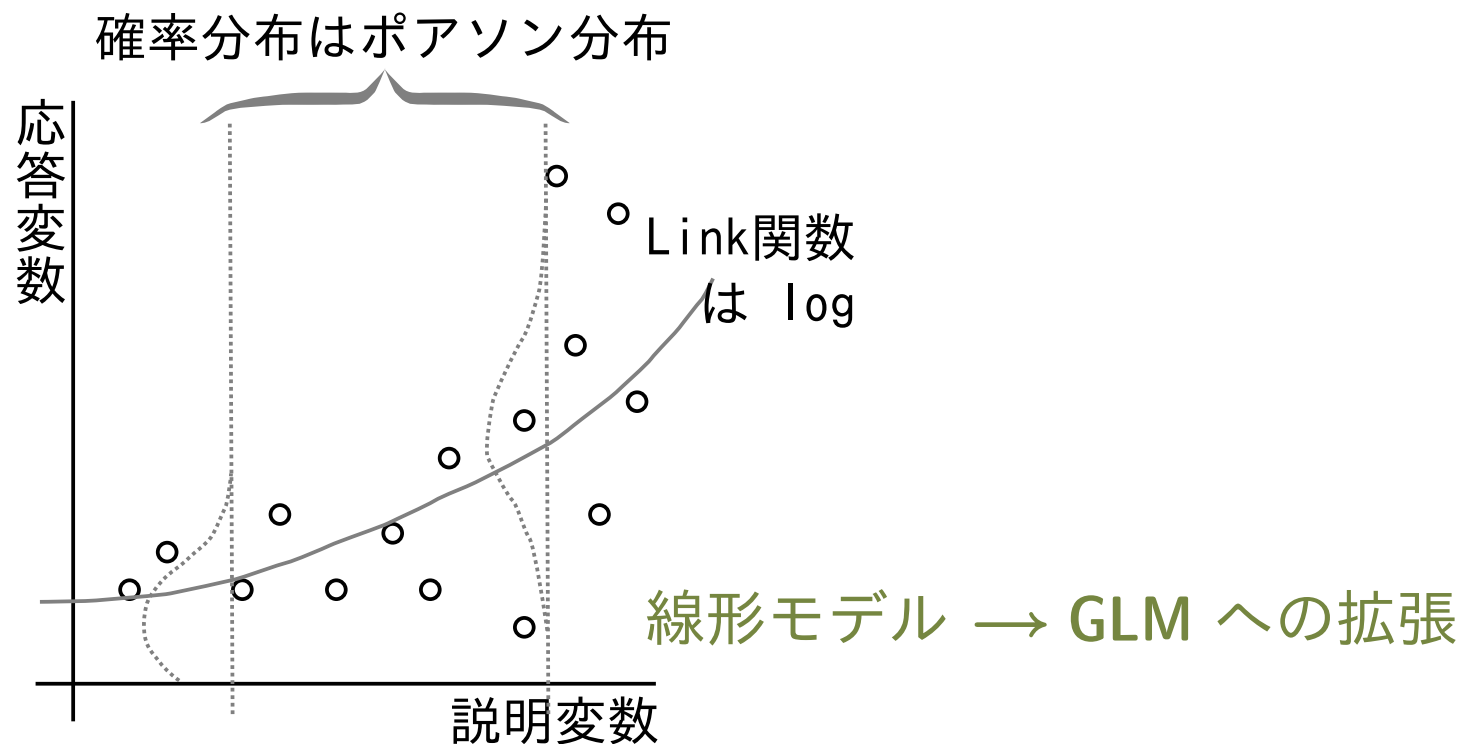
$$\begin{aligned} \text{(応答変数の平均)} &= \text{定数 (切片)} \\ &+ \text{(係数 1)} \times \text{(説明変数 1)} \\ &+ \text{(係数 2)} \times \text{(説明変数 2)} \\ &+ \text{(係数 3)} \times \text{(説明変数 3)} \\ &+ \dots \end{aligned}$$

(交互作用項については粕谷さんが説明してくれました)

線形モデルの発展



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

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 = $f(\text{線形予測子})$ という関係がある
- 線形予測子: $\beta_0 + \beta_1 x_1 + \beta_2 + \dots$, 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

さてさて、そろそろワリ算な統計学について……

世間でよくみかけるおススメできない作法の例

- ある調査地 i で N_i 本の樹木のうち k_i 本で開花していた
- 調査地 i の開花確率を $p_i = k_i/N_i$ とした
- 別の調査地 j の開花確率を $p_j = k_j/N_j$ とした
- 調査地 i と j の間で開花確率が異なるかどうか、 p_* が正規分布にしたがうと仮定して「ゆーい差を検定」した
- 確率は正規分布ではない、と指摘されたのでノンパラメトリック検定で「ゆーい差を検定」した

割算値ひねくるデータ解析はなぜよくないのか？

- 観測値 / 観測値がどんな確率分布にしたがうのか見とおしが悪く，さらに説明要因との対応づけが難しくなる
- 情報が失われる：「10 打数 3 安打」と「200 打数 60 安打」，「どちらも 3 割バッター」と言ってよいのか？
- 割算値を使わないほうが見とおしのよい，合理的なデータ解析ができる (今回の主題)
- したがって割算値を使ったデータ解析は不利な点ばかり，そんなことをする必要はどこにもない

避けられるわりざん，避けにくいわりざん

- 避けられる割算値

- 密度などの指数

例: 人口密度, specific leaf area (SLA) など

対策: **offset** 項わざ

- 確率

例: N 個のうち k 個にある事象が発生する確率

対策: ロジスティック回帰など**二項分布モデル**で

- 避けにくい割算値

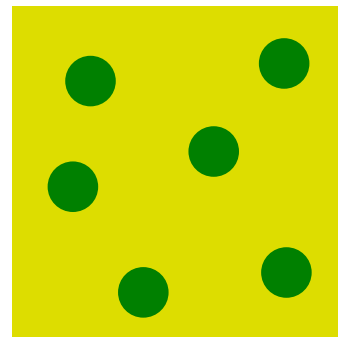
- 測定機器が内部で割算した値を出力する場合
- 割算値で作図せざるをえない場合があるかも

2. 「脱」割算の offset 項わざ

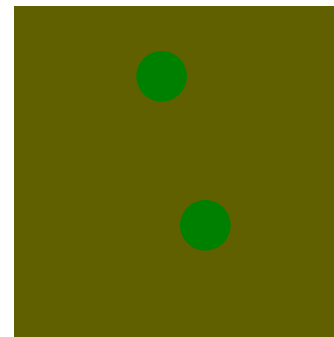
ポアソン回帰を強めてみる

今日の例題 B: 調査区画内の個体数は明るさで変わるか?

- 何か架空の植物個体の数が「明るさ」 x に応じてどう変わるかを知りたい
- 明るさは $\{0.1, 0.2, \dots, 1.0\}$ の 10 段階で観測した



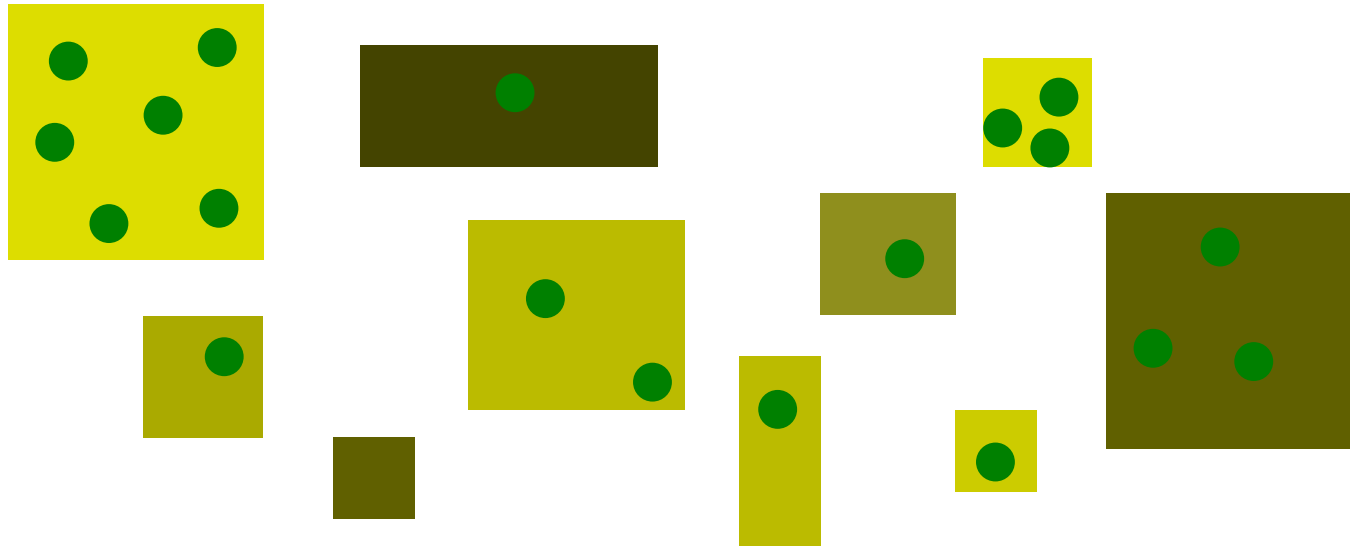
x 大
明るい



x 小
暗い

これだけなら単純に `glm(..., family = poisson)`
すればよいのだが……

「場所によって調査区の面積を変えました」?!!



- 明るさ x と面積 A を同時に考慮する必要あり
- ただし「密度 = 個体数 / 面積」といった割算値解析はやらない!
- `glm()` の `offset` 項わざわざうまく対処できる
- ともあれその前に観測データを図にしてみる

R の data.frame: 面積 Area, 明るさ x, 個体数 y

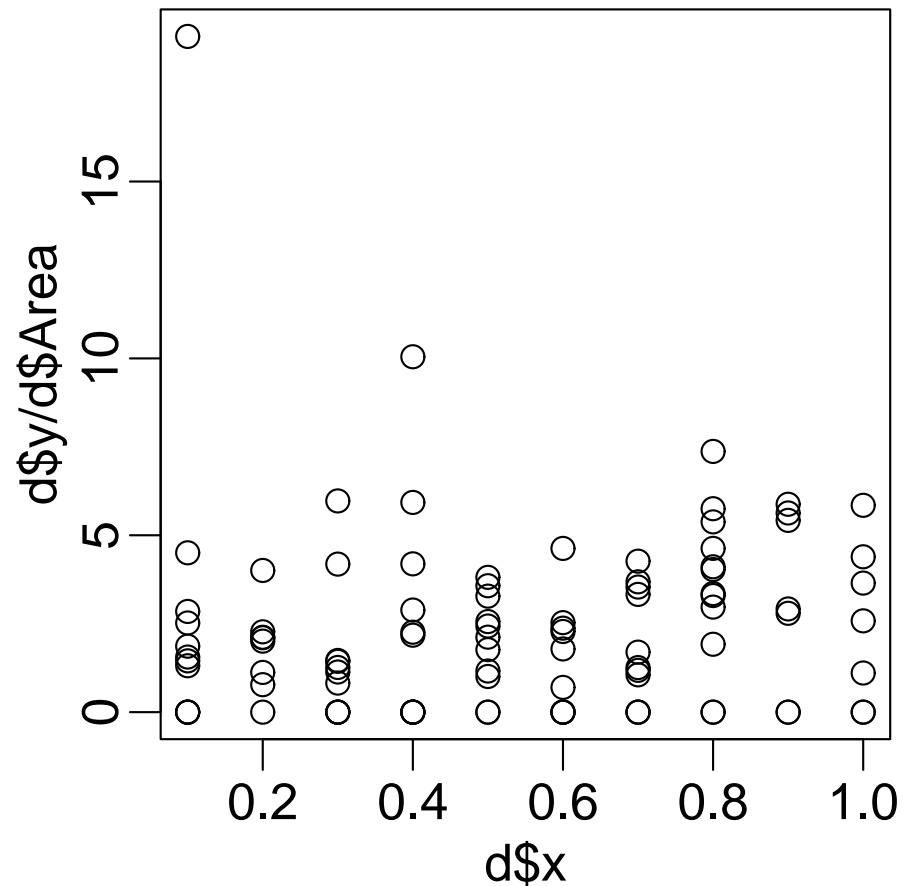
```
> load("d2.RData")
```

```
> head(d, 8) # 先頭 8 行の表示
```

	Area	x	y
1	0.017249	0.5	0
2	1.217732	0.3	1
3	0.208422	0.4	0
4	2.256265	0.1	0
5	0.794061	0.7	1
6	0.396763	0.1	1
7	1.428059	0.6	1
8	0.791420	0.3	1

ありがちな明るさ vs 割算値の図

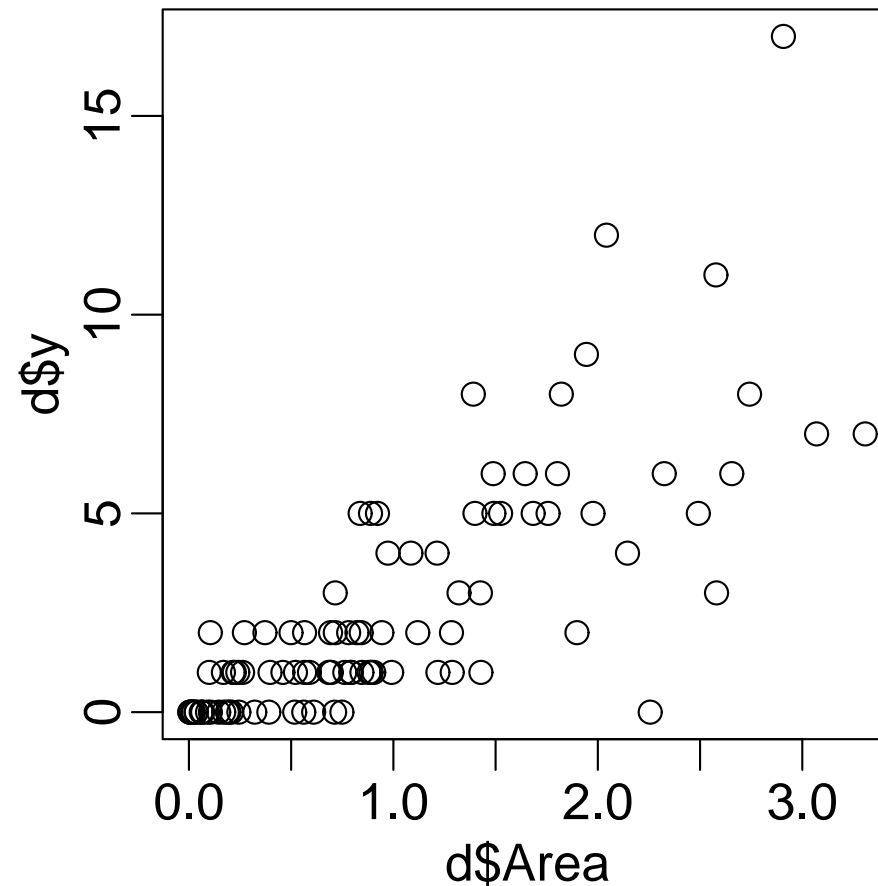
```
plot(d$x, d$y / d$Area)
```



- いまいちよくわからない……?

割算値ヤメて面積 A vs 個体数 y の図

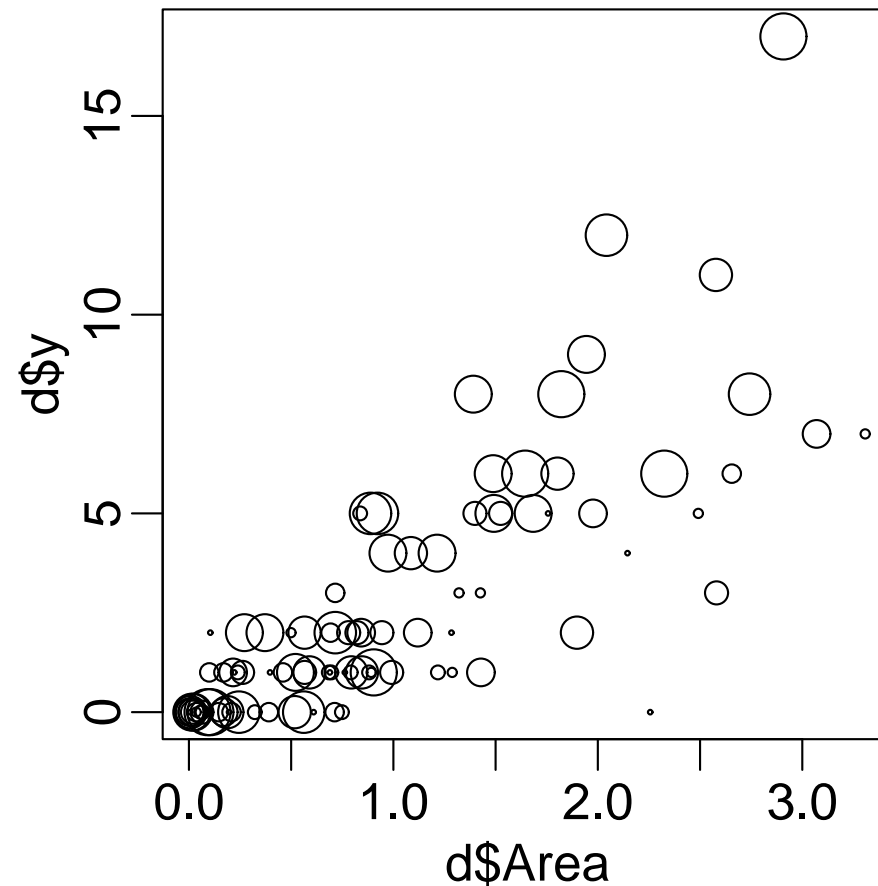
```
plot(d$Area, d$y)
```



- 面積 A とともに区画内の個体数 y が増大するようだ

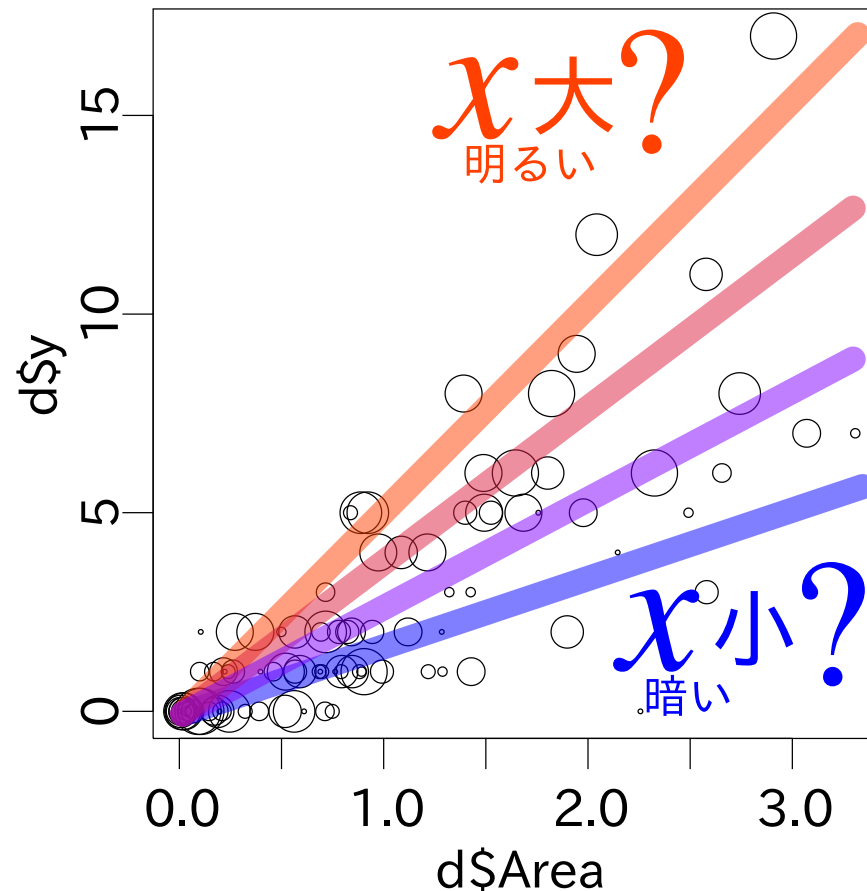
明るさ x の情報 (マルの大きさ) も図に追加

```
plot(d$Area, d$y, cex = d$x * 2)
```



- 同じ面積でも明るいほど個体数が多い?

密度が明るさ x に依存する統計モデル



- 区画内の個体数 y の平均は面積 \times 密度
- 密度は明るさ x で変化する
- ……といった統計モデルを作りたい!

「平均個体数 = 面積 × 密度」モデル

1. ある区画 i の応答変数 y_i は平均 λ_i のポアソン分布にしたがうと仮定:

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

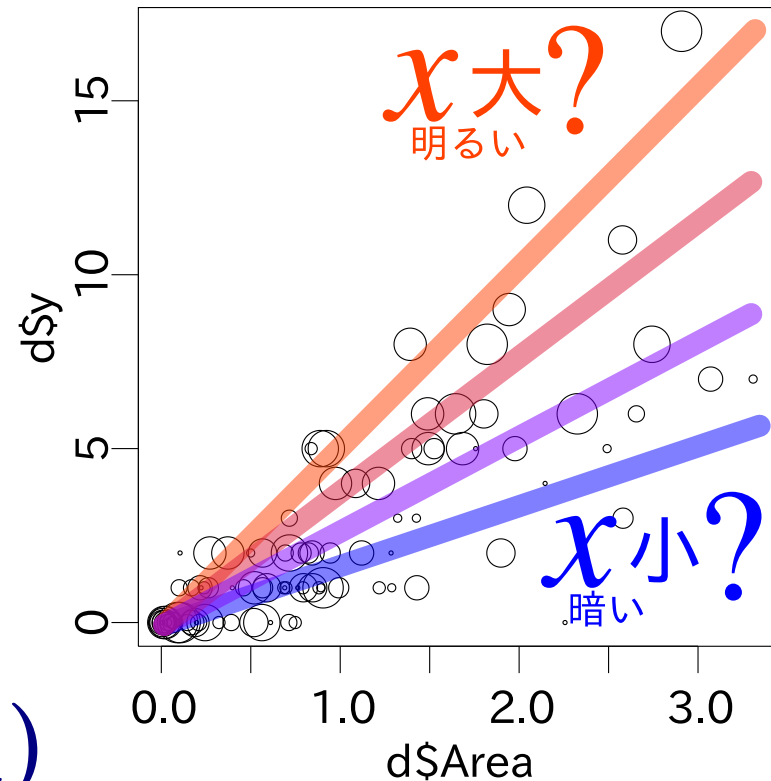
2. 平均値 λ_i は面積 A_i に比例し、密度は明るさ x_i に依存する

$$\lambda_i = A_i \exp(a + bx_i)$$

$$\lambda_i = \exp(a + bx_i + \log(A_i))$$

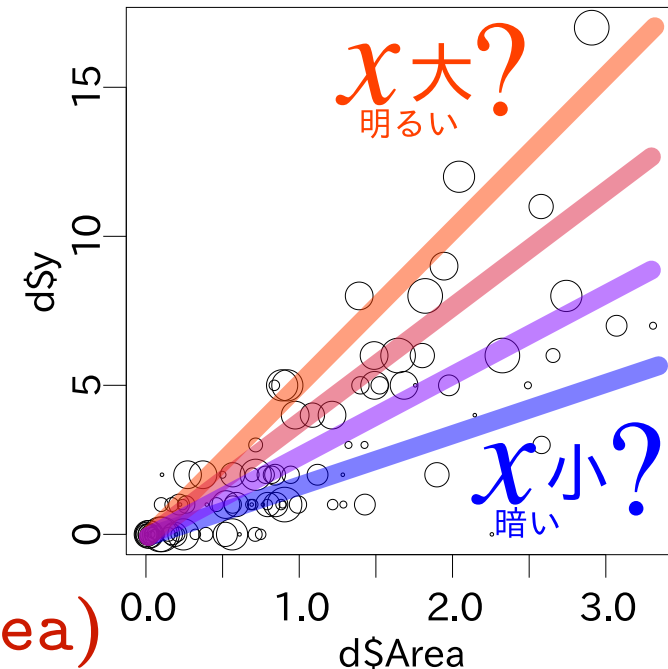
$$\log(\lambda_i) = a + bx_i + \log(A_i)$$

$\log(A_i)$ を offset 項とよぶ



この問題は GLM であつかえる! offset 項ワザ

- family: poisson, ポアソン分布
- link 関数: "log"
- モデル式: $y \sim x$
- offset 項の指定: $\log(\text{Area})$



- 線形予測子 $z = a + b x + \log(\text{Area})$
 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 $\lambda = \exp(z) = \exp(a + b x + \log(\text{Area}))$
- 応答変数は平均 λ のポアソン分布に従う: $y \sim \text{Pois}(\lambda)$

glm() 関数の指定

```
fit <- glm(  
  y ~ x,  
  family = poisson(link = "log")  
  data = d,  
  offset = log(Area)  
)
```

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

offset の指定

リンク関数の指定 (省略可)

R の glm() 関数による推定結果

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

Call:

```
glm(formula = y ~ x, family = poisson(link = "log"), data = d,  
  offset = log(Area))
```

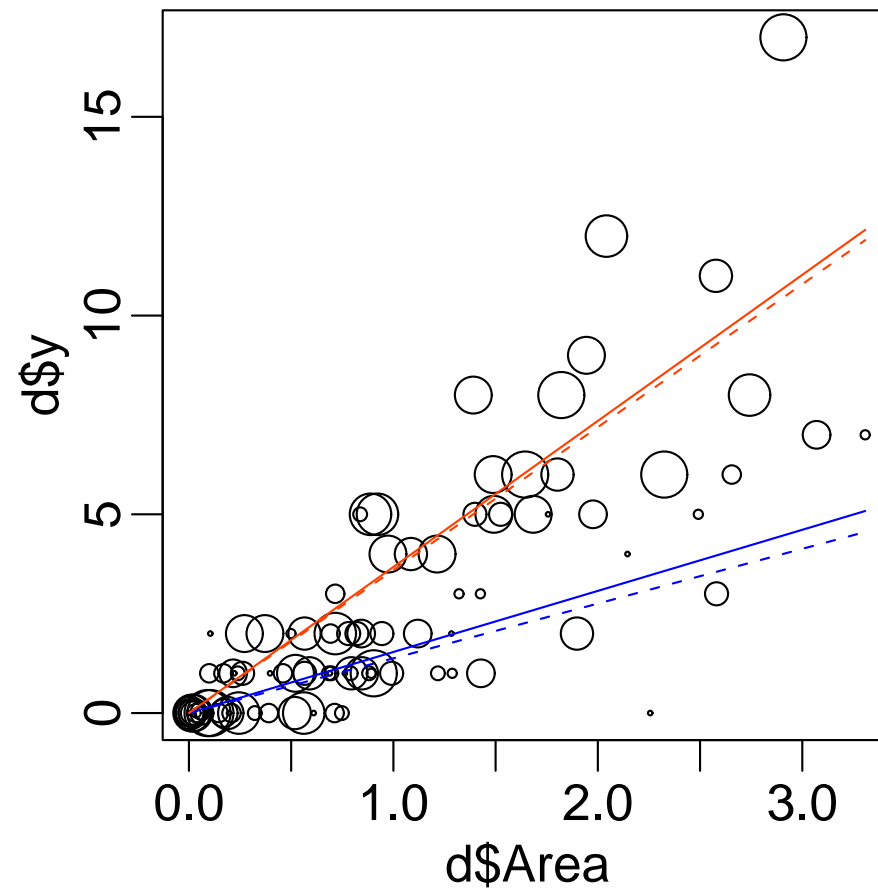
(... 略...)

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.321	0.160	2.01	0.044
x	1.090	0.227	4.80	1.6e-06

Coefficients は説明変数の係数という意味

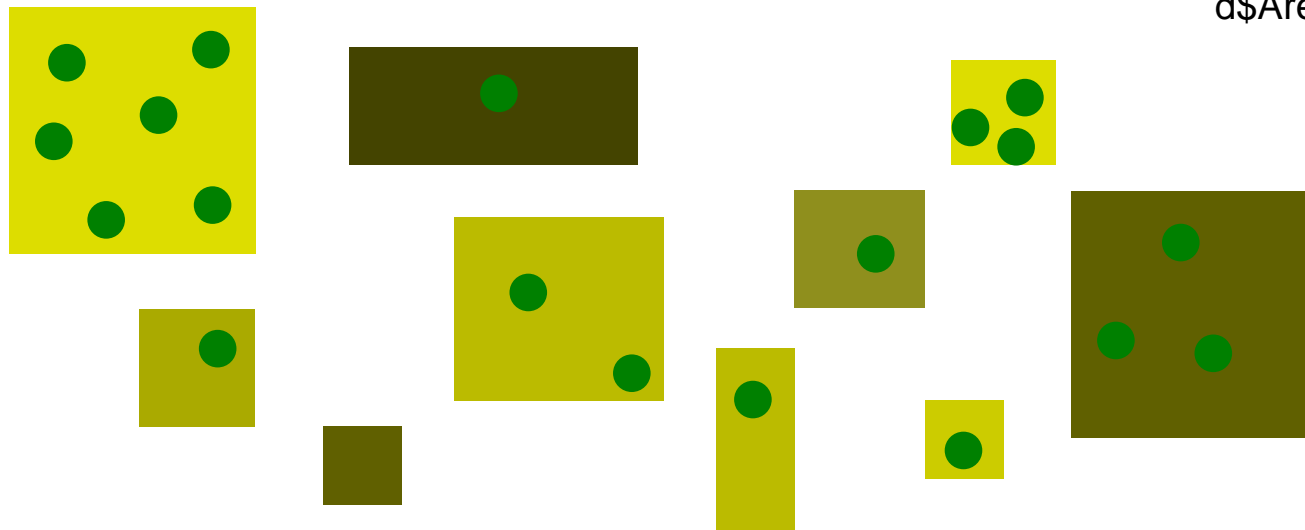
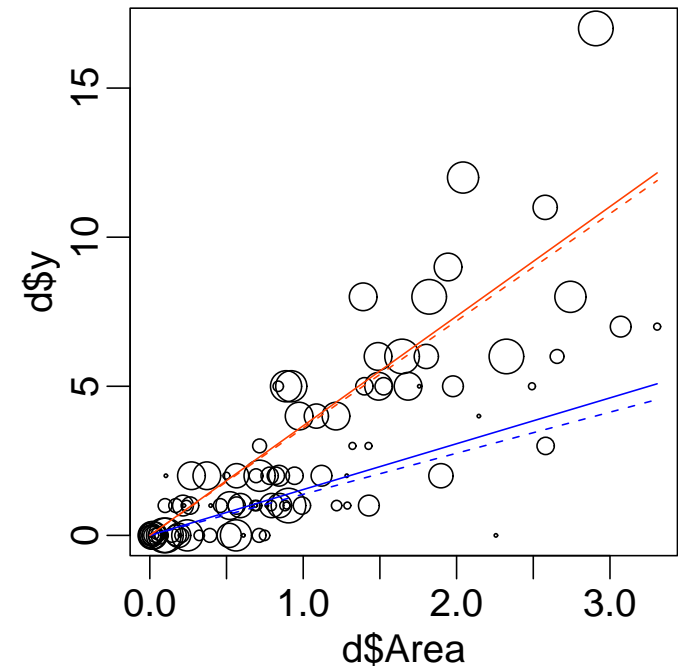
推定結果を図にしてみる



- 赤は明るさ $x = 0.9$, 青は $x = 0.1$
- 実線は `glm()` の推定結果, 破線はデータ生成時に指定した関係

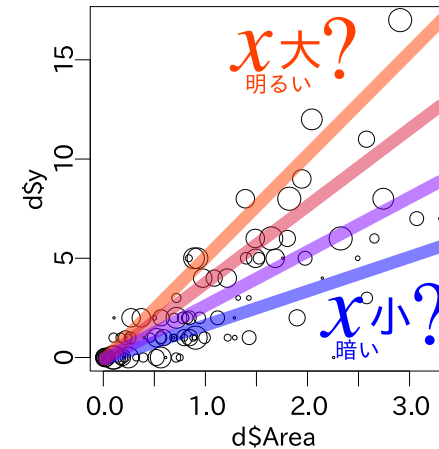
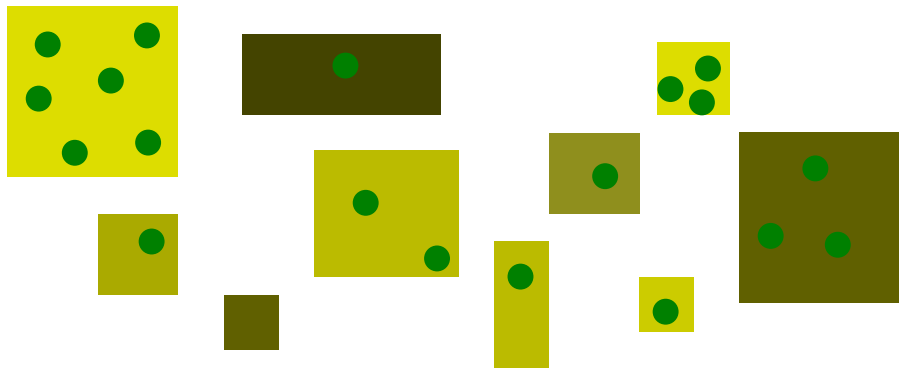
まとめ: glm() の offset 項わざで「脱」割算

- 平均値が面積などに比例する場合は、この面積などを **offset 項** として指定する
- 平均 = 面積 × 密度, というモデルの **密度** を $\exp(\text{線形予測子})$ として定式化する



今日のハナシ: 久保後半のまとめ

1. データ解析は統計モデリングだ
2. 割算するとわけわからなくなる
3. offset 項わざで「密度」をあつかう



今日ハナさなかつた GLM 関連事項

- ロジスティック回帰
- `summary(glm(...))` したときに表示される Wald 統計量と Pr の解釈
- 最尤推定法と deviance
- AIC によるモデル選択
- 過分散 (overdispersion) と GLMM への道

……などなど……

<http://hosho.ees.hokudai.ac.jp/~kubo/ce/>

参照してください

GLM よくある質問 (1) 「確率分布わからん」

どうやって確率分布を選べばいいんですか？

応答変数のタイプに注目して選んでください

- $y = 0, 1, 2, 3, \dots$ (y の上限不明) ならポアソン分布
(`family = poisson`)
- $y = \{0, 1\}$, $y = \{0, 1, 2, \dots, N\}$ なら二項分布
(`family = binomial`)
- 連続かつ正值ならガンマ分布 (`family = Gamma`)
- それ以外の連続値なら正規分布 (`family = gaussian`)

R で一般化線形モデル: glm() 関数

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

- `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

GLM よくある質問 (2) 「もっとヘンな分布を！」

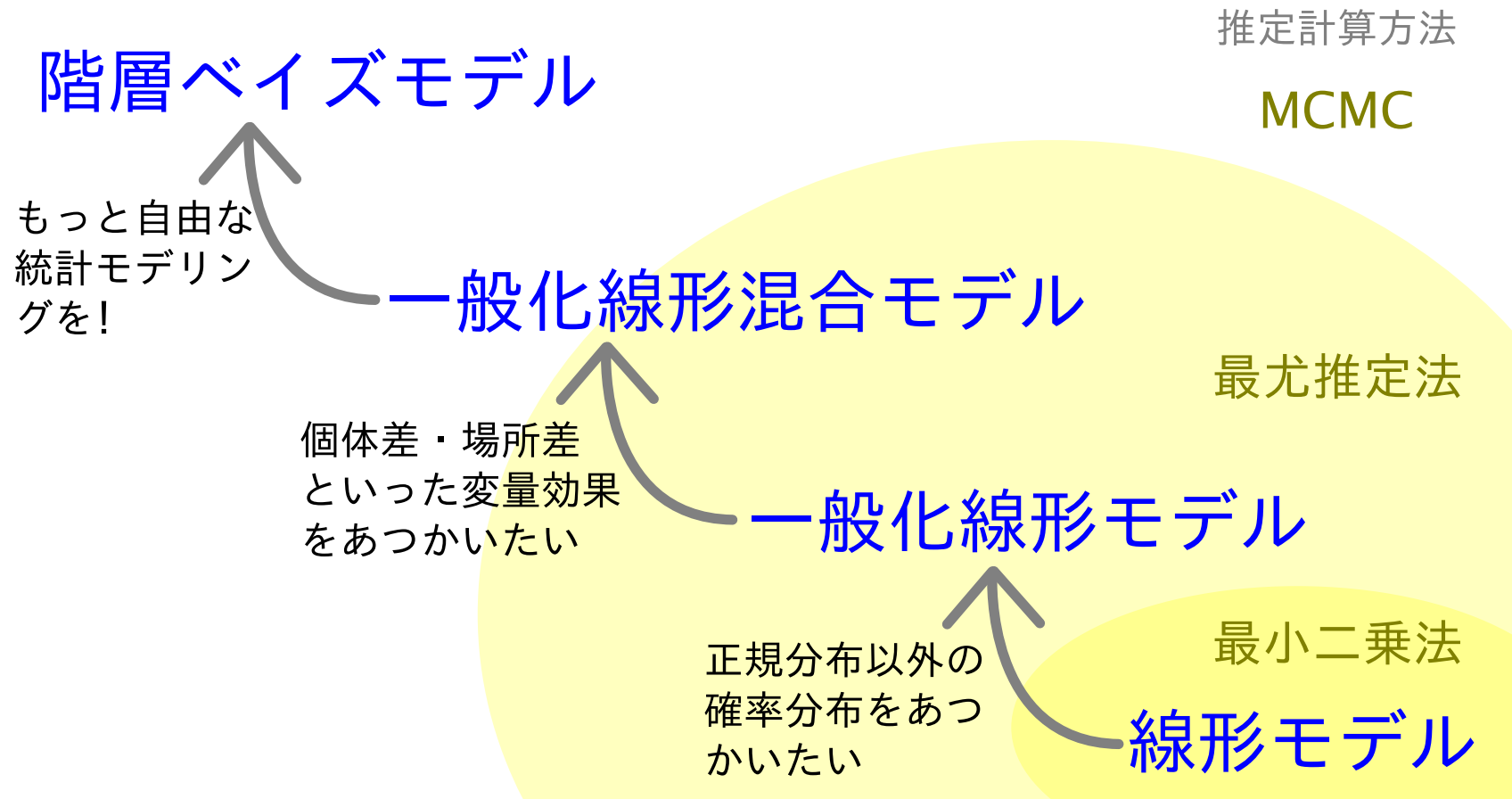
私のデータの確率分布はもっとヘンなんです！

GLMM や階層ベイズモデルに「ぱわーあっぷ」だ！

- まず、先ほどあげた「えらびかた」が基本です
- **GLMM/階層ベイズモデル**はこれらの**基本的な確率分布**を「**混ぜる**」ことでより複雑な状況に対処します
- 「混ぜる」ポイントは**個体差・場所差**といった **random effects** のモデリングです
- 「ぱわーあっぷ」にそなえて **GLM** の**基本**をよく勉強しましょう

おしまい

線形モデルの発展



<http://hosho.ees.hokudai.ac.jp/~kubo/ce/>

いろいろダウンロードできます