統計モデリング入門 2017 (j)

分割の統計モデル Categorical Data Analysis

久保拓弥 kubo@ees.hokudai.ac.jp

北大環境科学院の講義 http://goo.gl/76c4i

2017–11–15

ファイル更新時刻: 2017–11–08 17:11

ちょっと批判してみたい，よくみかける「お作法」

	Spc								
x	A	B	C	D	E	F	G	H	I
0	62	21	14	11	10	10	2	0	2
1	48	34	22	17	16	7	2	1	1

こういう表をみたときに反射的に……

- 表の検定 だからカイ二乗検定やればいいやー
  - 「なんでも検定」かよー
  - データ解析 $\neq$ 検定 !! 検定はデータ解析の一部 !!
- 「5 以下のセル」があるから検定できーん
  - その根拠は何 ?!
- データを捨てれば検定できる– !!
  - 捨てるな !!

分割表であれ，どんなデータであれ

- まず「どんな統計モデルで説明できるか」を考える
- カウントデータの場合は，とりあえず GLM で説明できないか考えてみる
- 次の項目をきちんと区別しよう
  - データを発生させうる統計モデル
    (例: GLM や階層ベイズモデル)
  - 統計モデルのパラメーター推定方法
    (例: 最尤推定法や MCMC 法)
  - 推定結果の比較方法
    (例: Neyman-Pearson な検定，モデル選択，信用区間)

2×2 の分割表

粕谷さんの話に登場した内容もくりかえしつつ

今日の例題: 調査区画に出現した植物の個体数

- 調査区画はふたつだけ　　反復ぐらいとれよ……
- まず，ふたつの調査区を「さら地」にした
- 片方の調査区で何らかの「処理」 (水やりとか?)
  - 無処理区 $(x = 0)$ と処理区 $(x = 1)$
- ある時点で出現した植物を，種 (Spc) ごとにカウント
- とりあえず，個体数の多かった A 種 と B 種について調べることにした — 個体数 $\{y_{\mathrm{A},0}, y_{\mathrm{B},0}, y_{\mathrm{A},1}, y_{\mathrm{B},1}\}$
- 知りたいこと: 「処理」によって A 種・B 種の割合は変わるのか?

  え一……「割合」だけ?

R で分割表をあつかう

data.frame() と xtabs()

データを CSV ファイルとして出力

	A	B	C
1	y	x	Spc
2	286	0	A
3	85	0	B
4	378	1	A
5	148	1	B
6			

- 「CSV」として保存 (脱 ゑくせる!)
- d2.csv というファイル名にする
- d2.csv の内容

```
y,x,Spc
286,0,A
85,0,B
378,1,A
148,1,B
```

データを R によみこみ，data.frame に変換

```
> d2 <- read.csv("d2.csv") # よみこんで，data.fram
変換!
> d2  # d2 という data.frame を表示
   y x Spc
1 286 0   A
2  85 0   B
4 378 1   A
5 148 1   B
```

xtabs: 分割表をあつかう R のクラス

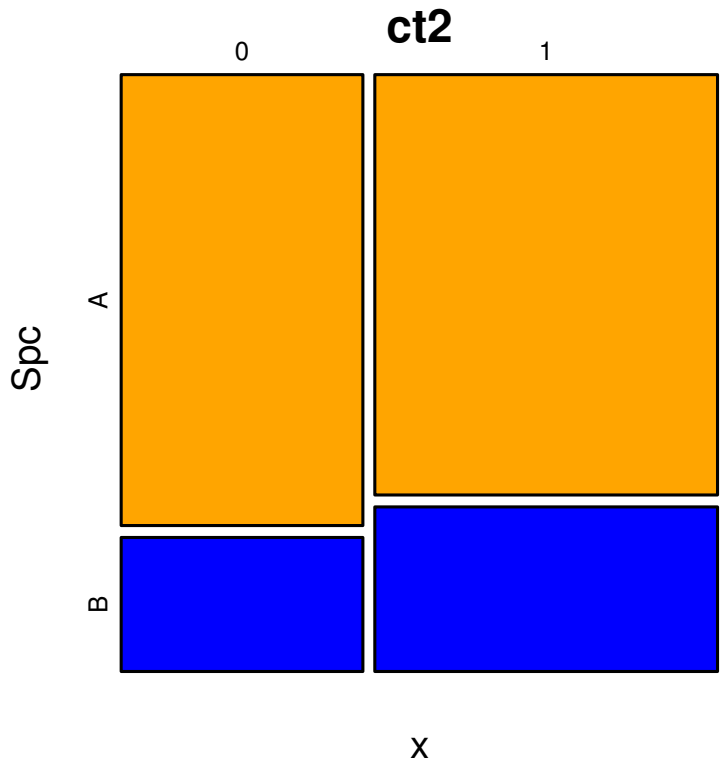
```
    y x Spc
1 286 0   A
2  85 0   B
4 378 1   A
5 148 1   B

> (ct2 <- xtabs(y ~ x + Spc, data = d2))
   Spc
x     A   B
  0 286  85
  1 378 148
```

xtabs: 自由自在に集計できる

```
> xtabs(y ~ x, data = d2)
x
  0   1
371 526
> xtabs(y ~ Spc, data = d2)
Spc
  A   B
664 233
> xtabs(y ~ Spc + x, data = d2)
    x
Spc   0   1
  A 286 378
  B  85 148
```

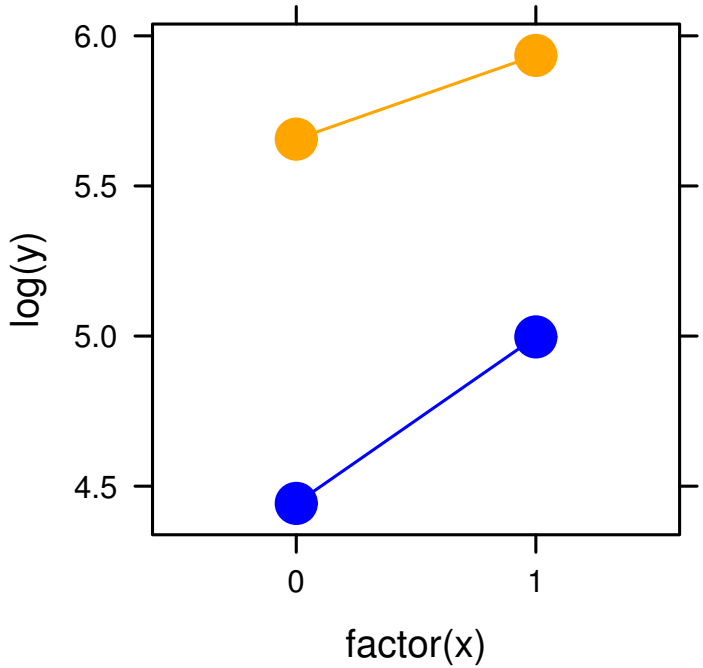
xtabs: 分割表の図示

```
    Spc
x     A   B
  0 286  85
  1 378 148
> plot(ct2, col = c("orange", "blue"))
```



library(lattice) を使った図示

```
   Spc
x     A   B
  0 286  85
  1 378 148
> library(lattice)
> xyplot(log(y) ~ factor(x), data = d2, groups = Spc, type = "b")
```



2×2 分割表の統計モデル

まずは二項分布の GLM から

ロジスティック回帰 logistic regression

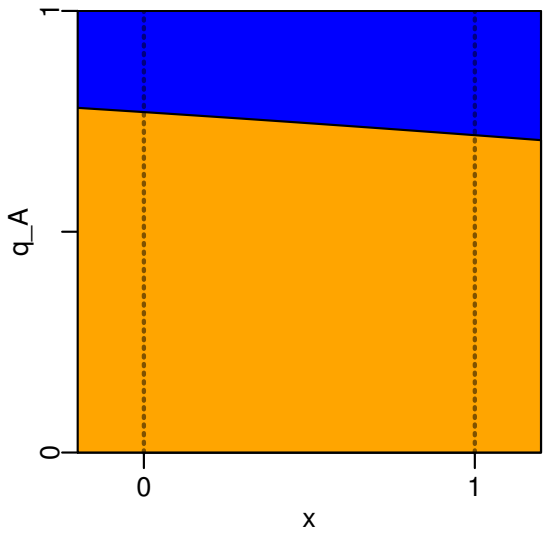
二項分布の GLM を適用してみる

$$
\begin{aligned}
y_{\mathrm{A},x} &\sim \mathrm{Binom}(q_{\mathrm{A},x},\ y_{\mathrm{A},x} + y_{\mathrm{B},x}) \\
\mathrm{logit}(q_{\mathrm{A},x}) &= a_{\mathrm{A}} + b_{\mathrm{A}} x
\end{aligned}
$$

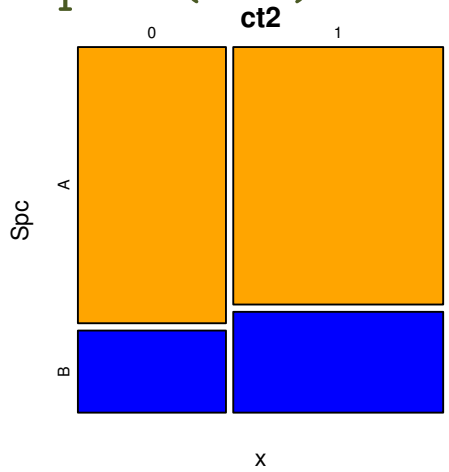
```
   Spc
x     A   B
  0 286  85
  1 378 148
> summary(glm(ct2 ~ c(0, 1), data = d2, family = binomial))
(... 略...)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.213      0.124    9.82   <2e-16
c(0, 1)       -0.276      0.157   -1.76    0.079
```

ロジスティック回帰の推定にもとづく予測

$$\mathrm{logit}(q_{\mathrm{A},x}) = 1.213 + (-0.276)x$$



`plot(ct2)` によるデータ図示



	モデル	AIC
x に依存する	$\mathrm{logit}(q_{\mathrm{A},x}) = a_{\mathrm{A}} + b_{\mathrm{B}}x$	16.5
x に依存しない	$\mathrm{logit}(q_{\mathrm{A},x}) = a_{\mathrm{A}}$	17.6

2×2 分割表の統計モデル

次にポアソン分布の GLM であつかってみる

とりあえず「分割方式」で

ポアソン回帰 Poisson regression

対数線形モデル log-linear model

ポアソン分布の GLM (分割方式) — A 種

$$
\begin{aligned}
y_{\mathrm{A},x} &\sim \mathrm{Pois}(\lambda_{\mathrm{A},x}) \\
\log(\lambda_{\mathrm{A},x}) &= \alpha_{\mathrm{A}} + \beta_{\mathrm{A}} x
\end{aligned}
$$

```
> # SpcA だけ
> summary(glm(y ~ x, data = d2[d2$Spc == "A",], family = poisson)
(... 略...)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.6560     0.0591   95.65  < 2e-16
x             0.2789     0.0784    3.56  0.00037
```

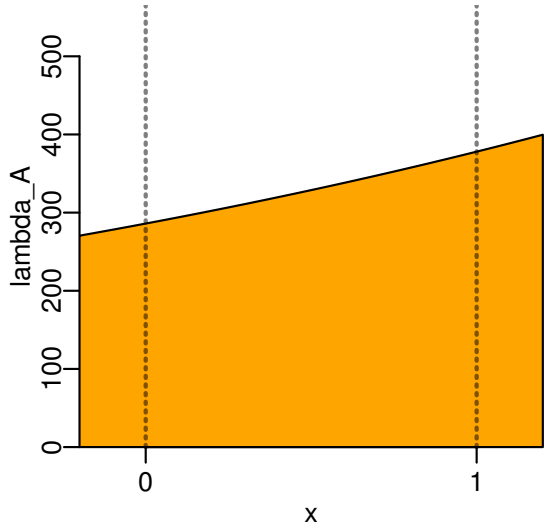
ポアソン分布の GLM (分割方式) — B 種

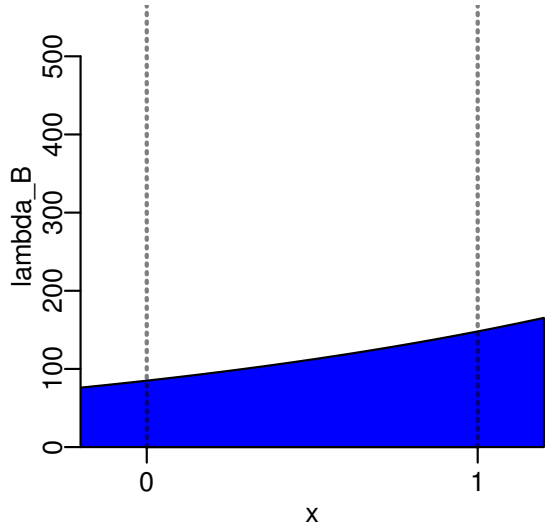
$$
\begin{aligned}
y_{\mathrm{B},x} &\sim \mathrm{Pois}(\lambda_{\mathrm{B},x}) \\
\log(\lambda_{\mathrm{B},x}) &= \alpha_{\mathrm{B}} + \beta_{\mathrm{B}} x
\end{aligned}
$$

```
> # SpcB だけ
> summary(glm(y ~ x, data = d2[d2$Spc == "B",], family = poisson)
(... 略...)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    4.443      0.108   40.96  < 2e-16
x              0.555      0.136    4.07  4.6e-05
```

ポアソン回帰の推定にもとづく予測

$$\log(\lambda_{A,x}) = 5.66 + 0.279x \qquad \log(\lambda_{B,x}) = 4.44 + 0.555x$$





モデル	AIC	モデル	AIC
$\lambda_{A,x} = \alpha_A + \beta_A x$	19.3	$\lambda_{B,x} = \alpha_B + \beta_B x$	17.1
$\lambda_{A,x} = \alpha_A$	30.1	$\lambda_{B,x} = \alpha_B$	32.4

ポアソン分布 GLM・二項分布 GLM のつながり

- 二項分布 GLM: $\mathrm{logit}(q_{A,x}) = a_A + b_A x$

$$q_{A,x} = \frac{1}{1 + \exp[-(a_A + b_A x)]}$$

- ポアソン分布: $\log(\lambda_{A,x}) = \alpha_A + \beta_A x$ など

$$\lambda_{A,x} = \exp(\alpha_A + \beta_A x)$$

$$\lambda_{B,x} = \exp(\alpha_B + \beta_B x)$$

……「A 種の割合」は?

$$\frac{\lambda_{A,x}}{\lambda_{A,x} + \lambda_{B,x}} = \frac{\exp(\alpha_A + \beta_A x)}{\exp(\alpha_A + \beta_A x) + \exp(\alpha_B + \beta_B x)}$$

$$= \frac{1}{1 + \exp[\alpha_B - \alpha_A + (\beta_B - \beta_A)x]}$$

係数の比較: ポアソン分布 GLM・二項分布 GLM のつながり

二項分布の GLM

$$q_{\mathrm{A},x} = \frac{1}{1 + \exp[-(a_{\mathrm{A}} + b_{\mathrm{A}} x)]}$$

ポアソン分布の GLM (分割方式)

$$\frac{\lambda_{\mathrm{A},x}}{\lambda_{\mathrm{A},x} + \lambda_{\mathrm{B},x}} = \frac{1}{1 + \exp[\alpha_{\mathrm{B}} - \alpha_{\mathrm{A}} + (\beta_{\mathrm{B}} - \beta_{\mathrm{A}})x]}$$

比較すると……

二項分布 GLM		ポアソン分布 GLM
a_{A}	$=$	$\alpha_{\mathrm{A}} - \alpha_{\mathrm{B}}$
b_{A}	$=$	$\beta_{\mathrm{A}} - \beta_{\mathrm{B}}$

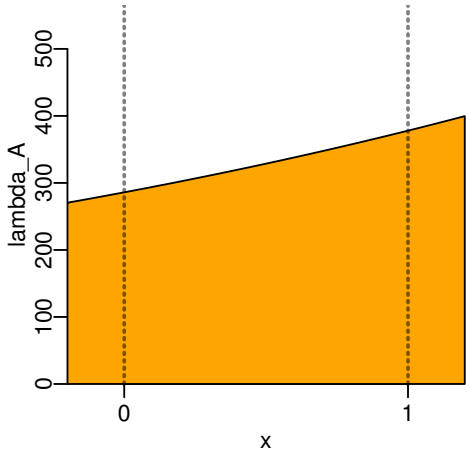
比較: 二項分布とポアソン分布の GLM

二項分布 GLM		ポアソン分布 GLM
$a_A = 1.213$	=	$\alpha_A - \alpha_B$
$b_A = -0.276$	=	$\beta_A - \beta_B$

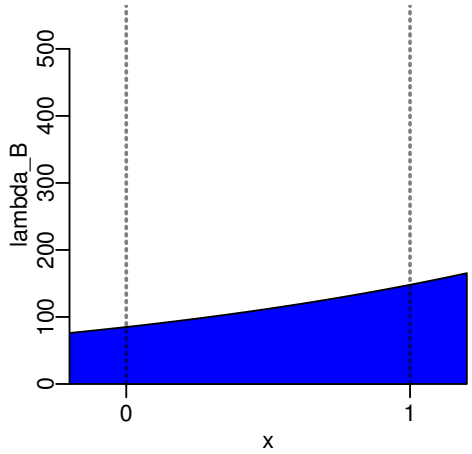
```
> 二項分布 GLM (A 種の比率)
> glm(ct2 ~ c(0, 1), data = d2, family = binomial)
(Intercept)      c(0, 1)
      1.213       -0.276
> ポアソン分布 GLM (A 種の比率)
> glm(y ~ x, data = d2[d2$Spc == "A",], family = poisson)
(Intercept)            x
      5.656        0.279
> ポアソン分布 GLM (B 種の比率)
> glm(y ~ x, data = d2[d2$Spc == "B",], family = poisson)
(Intercept)            x
      4.443        0.555
```

図解: ポアソン分布 GLM・二項分布 GLM のつながり

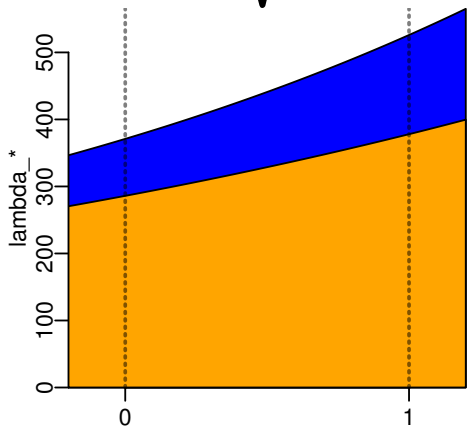
ポアソン分布の GLM (A 種)



ポアソン分布の GLM (B 種)

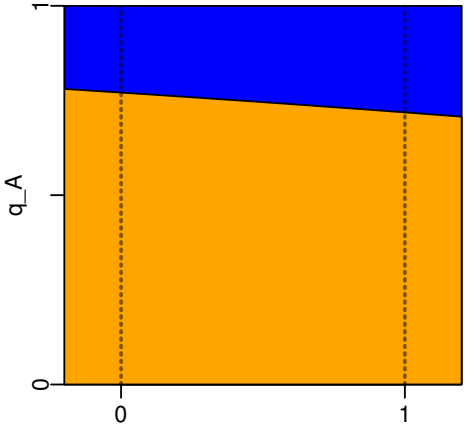


つみあげる



たいらに
押しつぶす

二項分布の GLM
(A 種+ B 種)



2×2 分割表の統計モデル

データを分割しないポアソン分布 GLM

「一括方式」 — こちらが便利かも?

ポアソン分布の GLM (一括方式)

交互作用項をうまく利用する

```
> summary(glm(y ~ x * Spc, data = d2, family = poisson))
(... 略...)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.6560     0.0591   95.65  < 2e-16
x             0.2789     0.0784    3.56  0.00037
SpcB         -1.2133     0.1235   -9.82  < 2e-16
x:SpcB        0.2757     0.1570    1.76  0.07921
(... 略...)
```

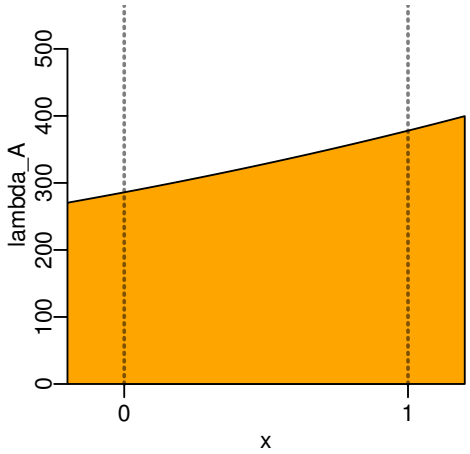
「分割方式」のポアソン分布 GLM と一致 $\rightarrow$ 二項分布 GLM とも一致

$$\alpha_{\mathrm{A}} = 5.66 \qquad \alpha_{\mathrm{B}} = 5.66 - 1.21$$

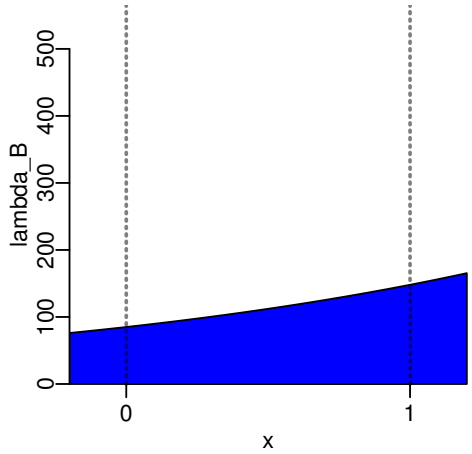
$$\beta_{\mathrm{A}} = 0.279 \qquad \beta_{\mathrm{B}} = 0.279 + 0.276$$

ポアソン・二項分布両 GLM のつながり (再)

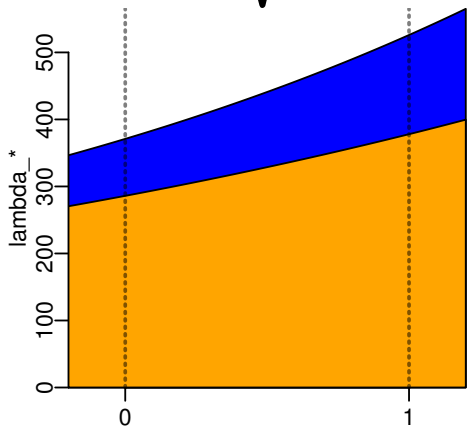
ポアソン分布の GLM (A 種)



ポアソン分布の GLM (B 種)

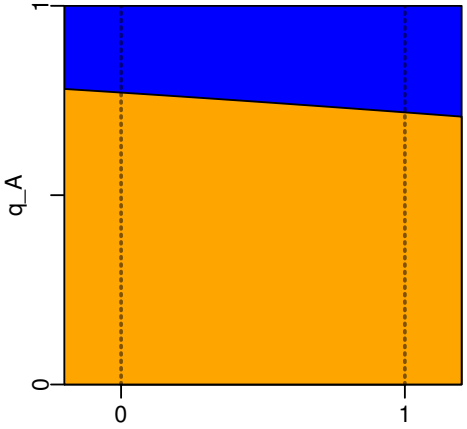


つみあげる



たいらに
押しつぶす

二項分布の GLM
(A 種+ B 種)



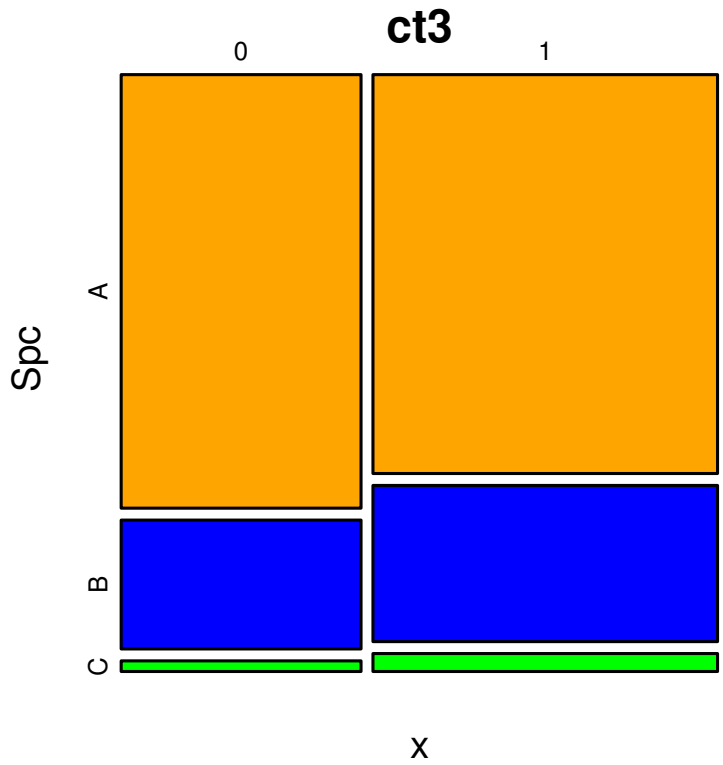
2×3 の分割表

多項分布の GLM か?

ポアソン分布の GLM か?

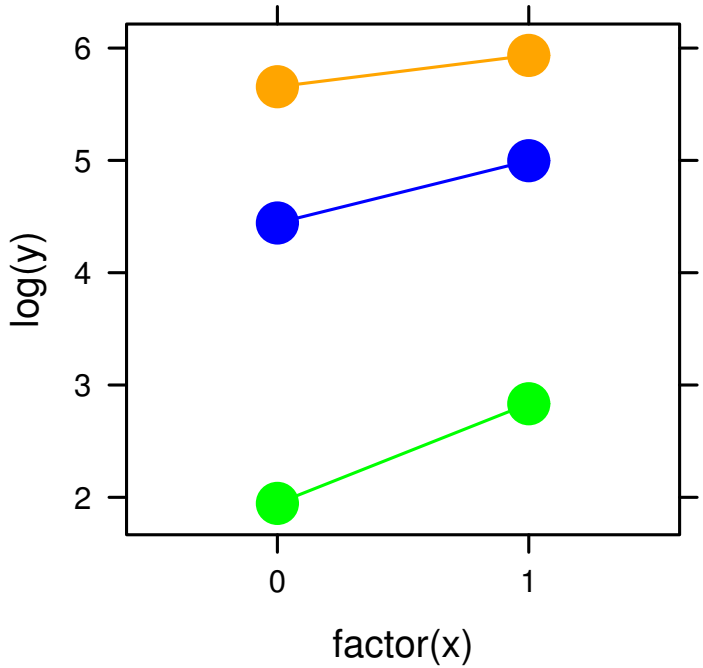
xtabs: 分割表の図示

```
    Spc
x      A   B   C
  0 286  85   7
  1 378 148  17
> plot(ct3, col = c("orange", "blue", "green"))
```



library(lattice) を使った図示

```
   Spc
x     A   B   C
  0 286  85   7
  1 378 148  17
> library(lattice)
> xyplot(log(y) ~ factor(x), data = d3, groups = Spc, type = "b")
```



ポアソン分布の GLM (一括方式)

```
> glm(y ~ x * Spc, data = d3, family = poisson)
(... 略...)
Coefficients:
(Intercept)           x        SpcB        SpcC     x:SpcB     x:SpcC
      5.656      0.279      -1.213      -3.710      0.276      0.608
```

「分割方式」のポアソン分布 GLM のパラメーターで言うと……

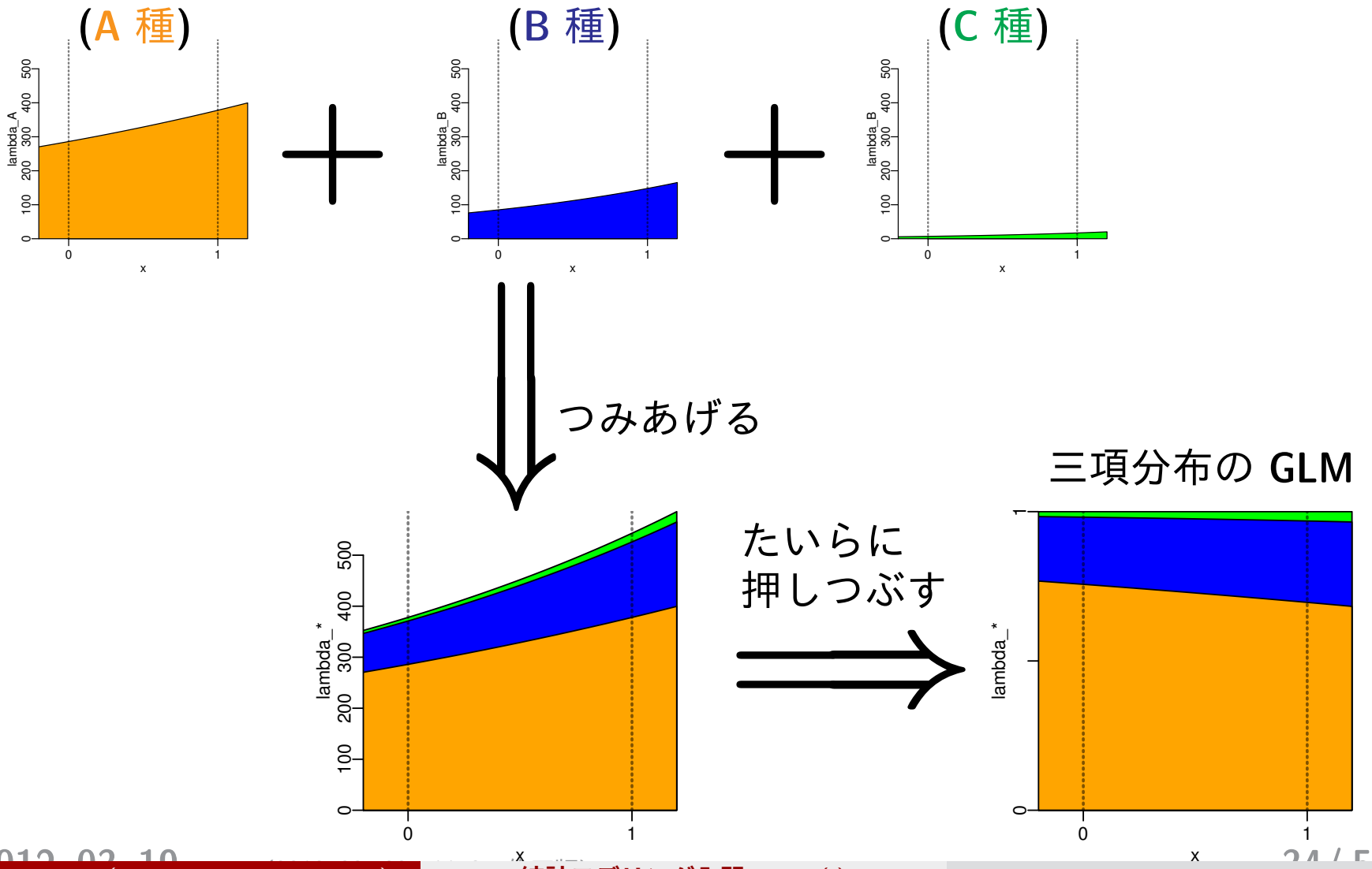
$$y_{A,x} \sim \mathrm{Pois}(\lambda_{A,x})$$
$$\log(\lambda_{A,x}) = \alpha_A + \beta_A x$$

$\alpha_A = 5.66$ $\alpha_B = 5.66 - 1.21$ $\alpha_C = 5.66 - 3.71$

$\beta_A = 0.279$ $\beta_B = 0.279 + 0.276$ $\beta_C = 0.279 + 0.608$

ポアソン分布 GLM の予測など

ポアソン分布の GLM



二項分布 GLM を拡張した多項分布 GLM

library(nnet) の multinom()

多項分布・ロジスティックな GLM

```
> ct3 # 分割表を表示
   Spc
x     A   B   C
  0 286  85   7
  1 378 148  17
> library(nnet) # nnet package よみこみ
> multinom(ct3 ~ c(0, 1))
(... 略...)
Coefficients:
  (Intercept) c(0, 1)
B     -1.2133 0.27552
C     -3.7097 0.60763
```

多項分布・ロジスティック GLM

$$
\begin{aligned}
y_{\mathrm{B},x} &\sim \text{Multinom}(q_{\mathrm{B},x},\ \text{3 種合計} \\
y_{\mathrm{C},x} &\sim \text{Multinom}(q_{\mathrm{C},x},\ \text{3 種合計} \\
\text{logit}(q_{\mathrm{B},x}) &= a_{\mathrm{B}} + b_{\mathrm{B}} x \\
\text{logit}(q_{\mathrm{C},x}) &= a_{\mathrm{C}} + b_{\mathrm{C}} x
\end{aligned}
$$

```
> # ポアソン分布 GLM と同じ推定値!
```

モデル選択はどうする?

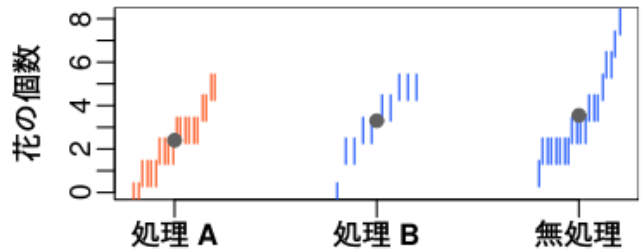
- すべての可能なくみあわせで調べるしかない
  - cf. 2004 年データ解析自由集会の久保発表

5. モデル選択基準が最良のものを採用する

すべてのグループわけについて AIC を計算する.

```
> results <- estimate.poisson(samples)
> cat(sapply(results, function(r)
  sprintf("Model %-12s, AIC = %.1f", r$tag, r$glm$aic)),
  sep  = "\n")
```

```
Model (A+B+C)     , AIC = 195.5
Model (A+B)(C)    , AIC = 194.7
Model (A+C)(B)    , AIC = 197.3
Model (A)(B+C)    , AIC = 192.9
Model (A)(B)(C)   , AIC = 194.8
```



AIC 最小のモデルを選ぶ

2×9 の分割表

……そろそろ GLM ではしんどくなってくる?

また別のデータ: 種数が 3 から 9 に増えた!

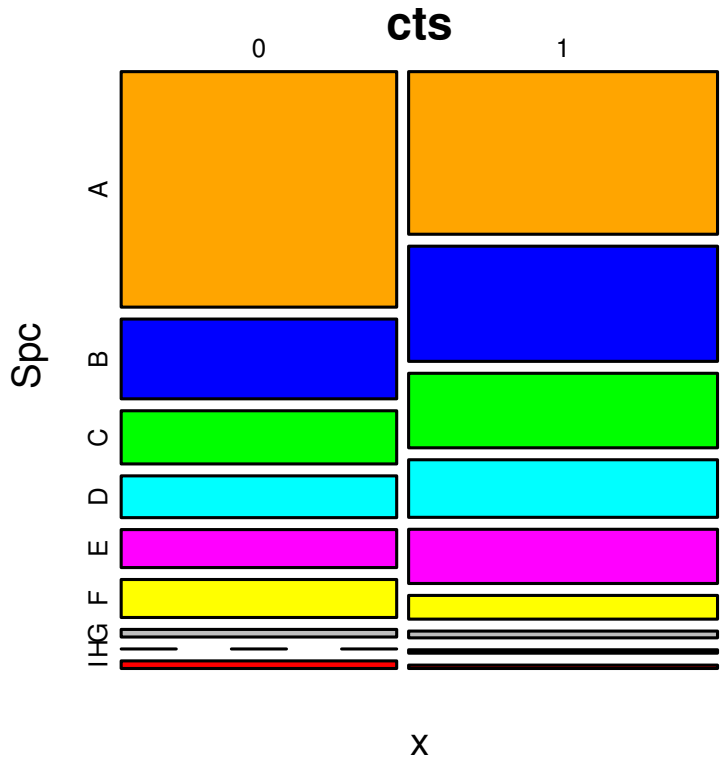
```
> d9
    y x Spc
1  62 0   A
2  21 0   B
3  14 0   C
4  11 0   D
5  10 0   E
6  10 0   F
7   2 0   G
8   0 0   H
9   2 0   I
10 48 1   A
(... 略...)
15  7 1   F
16  2 1   G
17  1 1   H
18  1 1   I
```

```
> ct9
   Spc
x    A  B  C  D  E  F  G  H  I
  0 62 21 14 11 10 10  2  0  2
  1 48 34 22 17 16  7  2  1  1
```

- 種ごとに個体数のばらつきがかなりある
- ゼロデータを含む

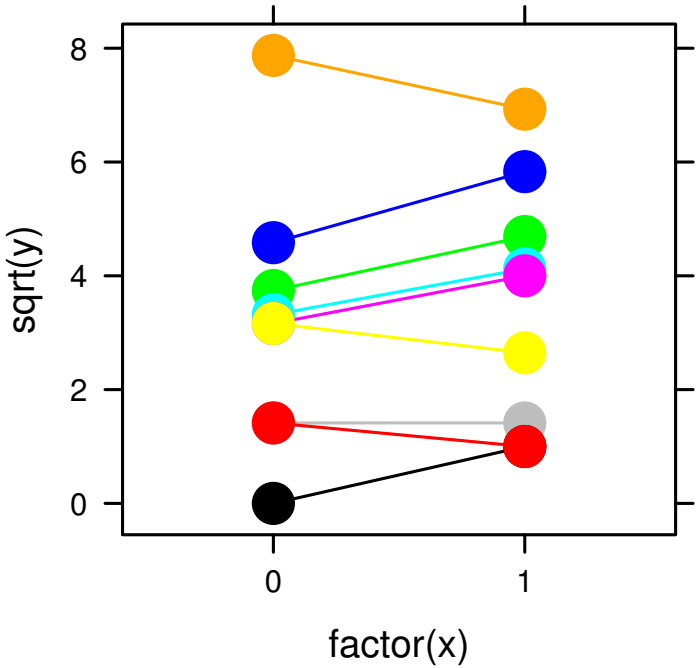
xtabs: 分割表の図示

```
    Spc
x     A  B  C  D  E  F  G  H  I
  0 62 21 14 11 10 10  2  0  2
  1 48 34 22 17 16  7  2  1  1
> plot(ct9, col = c(ごちゃごちゃと指定))
```



library(lattice) を使った図示

```
      Spc
   x    A  B  C  D  E  F  G  H  I
     0 62 21 14 11 10 10  2  0  2
     1 48 34 22 17 16  7  2  1  1
> library(lattice)
> xyplot(sqrt(y) ~ factor(x), data = d9, groups = Spc, type = "b")
```



ポアソン分布の GLM (一括方式)

```
> ct9
   Spc
x    A  B  C  D  E  F  G  H  I
  0 62 21 14 11 10 10  2  0  2
  1 48 34 22 17 16  7  2  1  1
> summary(glm(y ~ x * Spc, data = d9, family = poisson))
```

(Intercept)	x	SpcB	SpcC
4.127	-0.256	-1.083	-1.488
SpcD	SpcE	SpcF	SpcG
-1.729	-1.825	-1.825	-3.434
SpcH	SpcI	x:SpcB	x:SpcC
-26.430	-3.434	0.738	0.708
x:SpcD	x:SpcE	x:SpcF	x:SpcG
0.691	0.726	-0.101	0.256
x:SpcH	x:SpcI		
22.559	-0.437		

H 種 の推定値がかなりヘン!

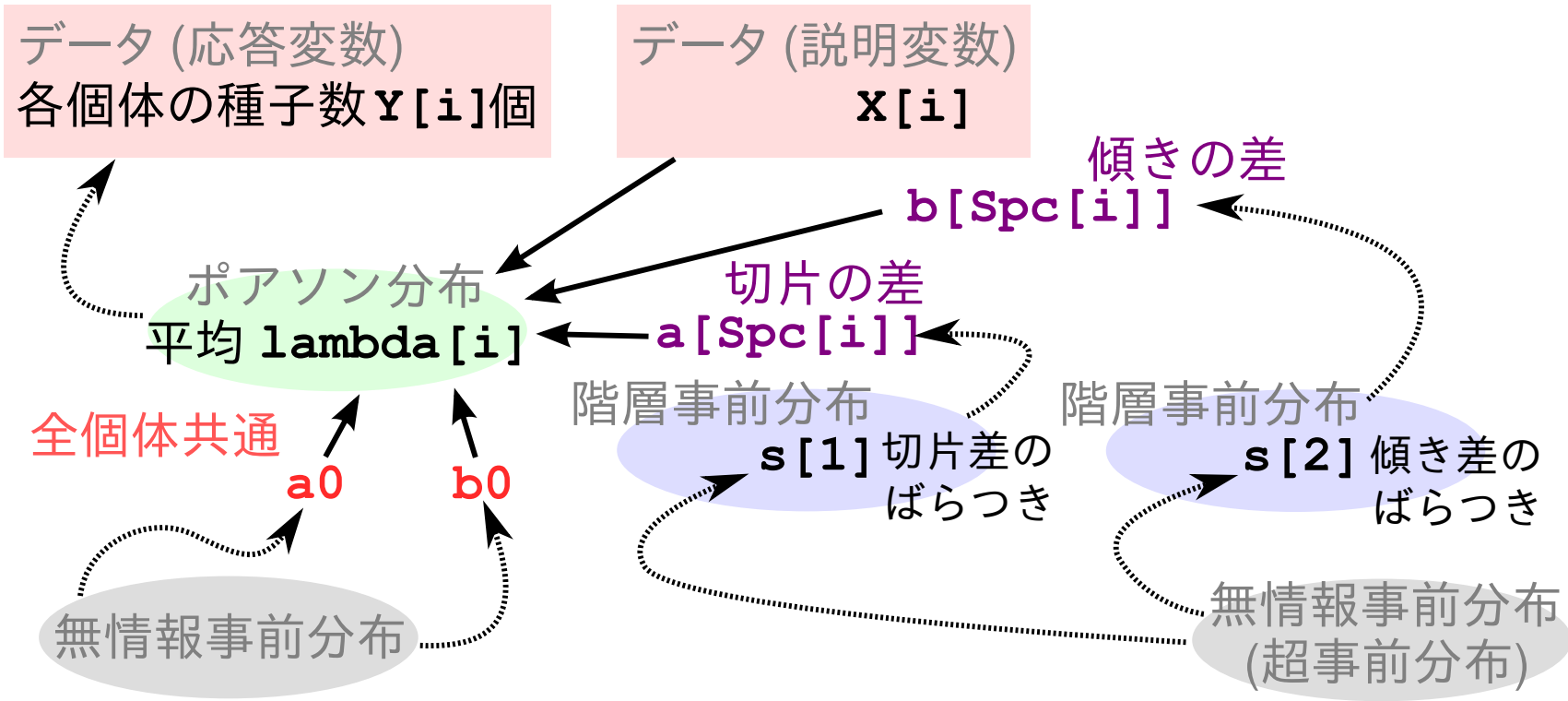
「なんでも glm()」方針の問題点

- 分割表がでかくなったときに，独立に推定されるパラメーター数がどんどん増えてしまう
- そのような場合，とくにゼロデータなどがあると，パラメーター推定が難しくなる
- モデル選択とかもしんどくなる

ポアソン分布の GLMM ならどうだろう?

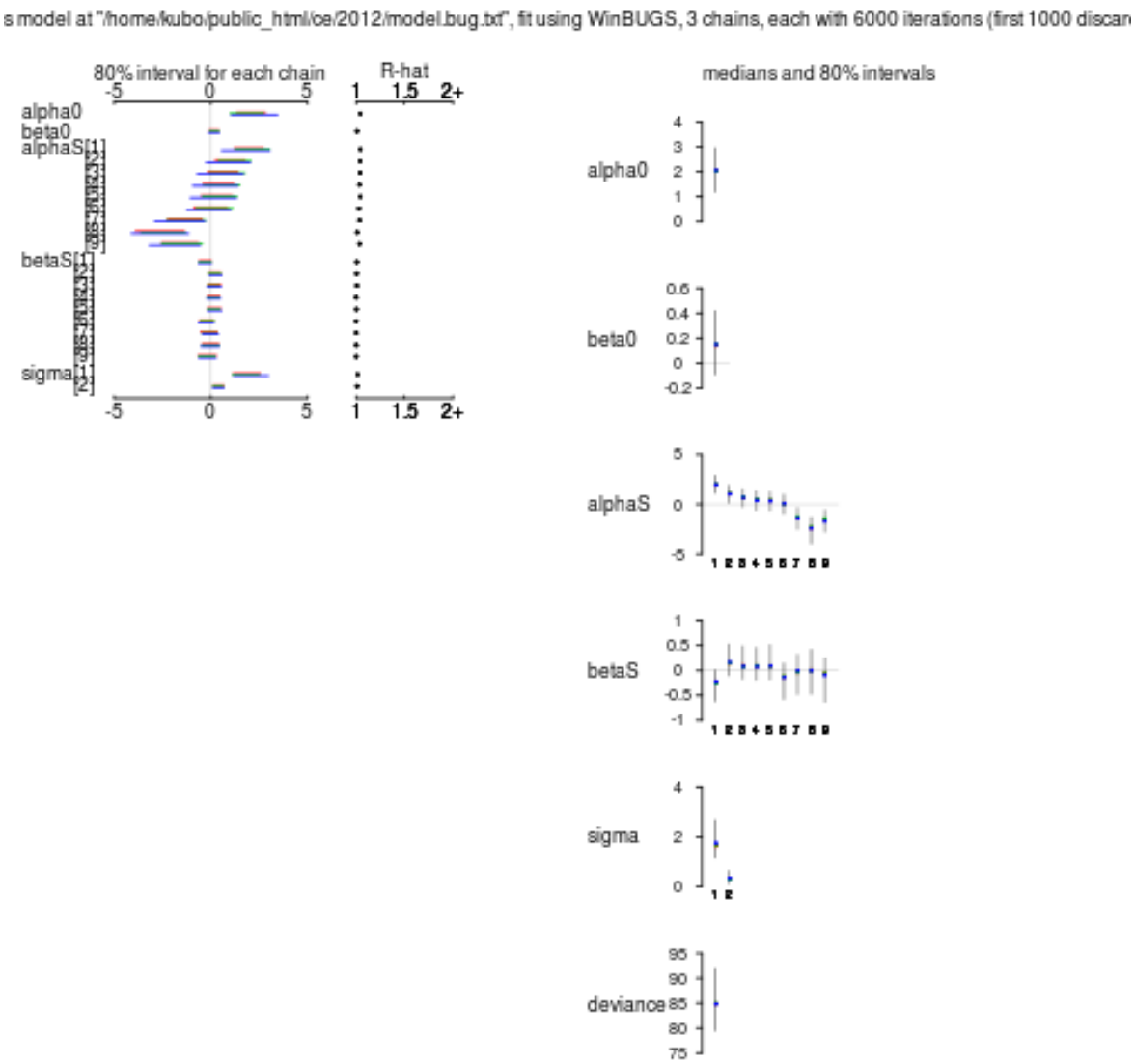
```
> (fit.glmm <- glmmML(y ~ x, data = d9, cluster = Spc, family = poisson))

             coef se(coef)     z Pr(>|z|)
(Intercept) 2.012    0.465 4.323  1.5e-05
x           0.114    0.120 0.956  3.4e-01

> fit.glmm$posterior.modes
[1]  1.926862  1.230935  0.807098  0.557225  0.483854
[6]  0.067305 -1.218522 -2.005846 -1.426968
```

分割表の階層ベイズモデル (線形ポアソン回帰)



- このようにデータを設計する
- WinBUGS を使ってパラメーター推定 (MCMC 法) をしたいので，BUGS コードで書く
- WinBUGS を R の下っぱとして使う

(時間があれば WinBUGS 実演)

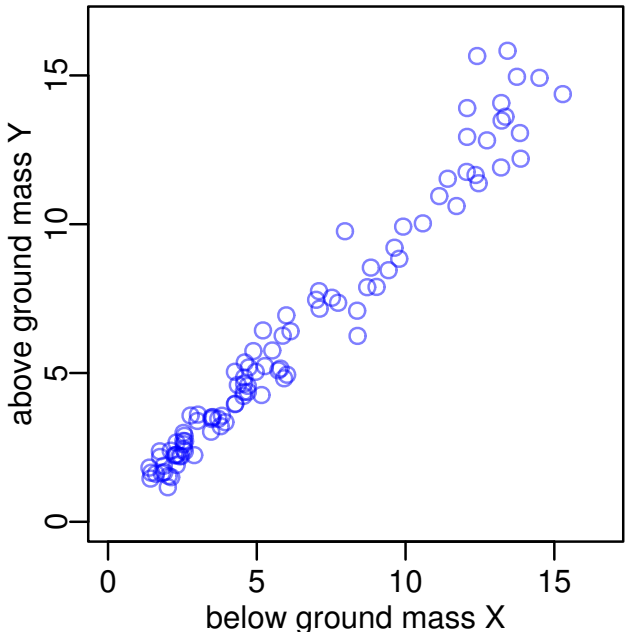


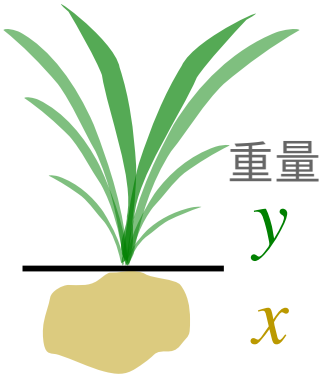
例題 1

「アロメトリーな回帰」はヤメて

重量分割モデルを作ってみよう

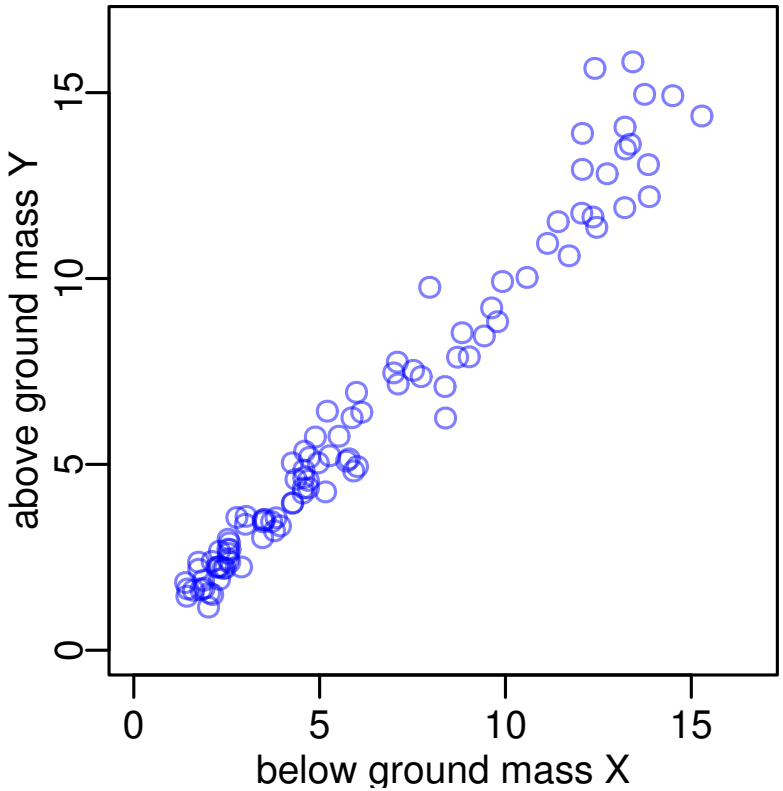
架空データ 1: 地下部・地上部の重量

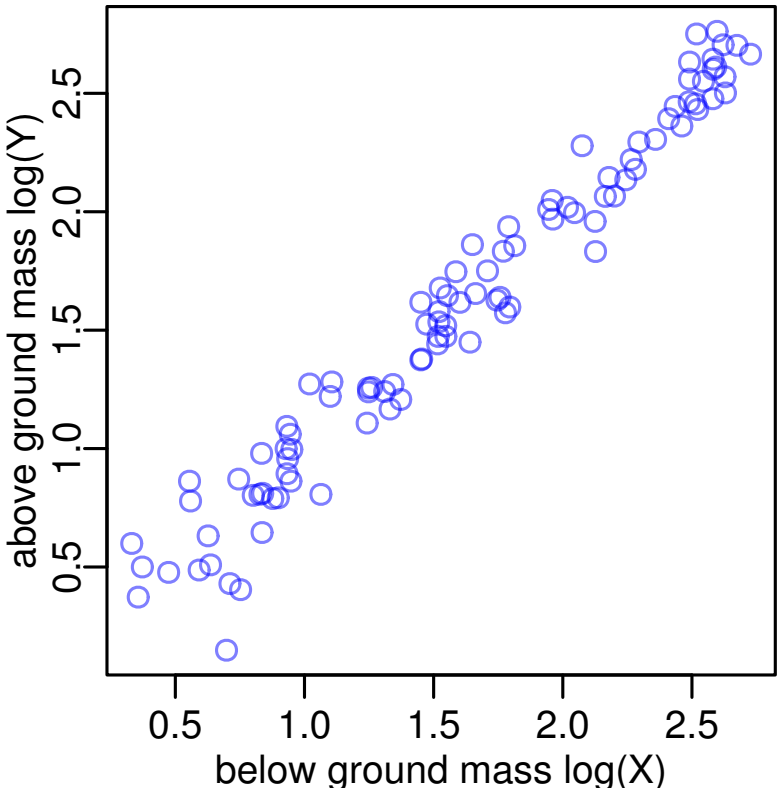




- 対数をとらなくても「直線」にのってる?
- つまり，むしろアイソメトリック (isometric)?
- 重量が重くなるとばらつきが大きくなる?

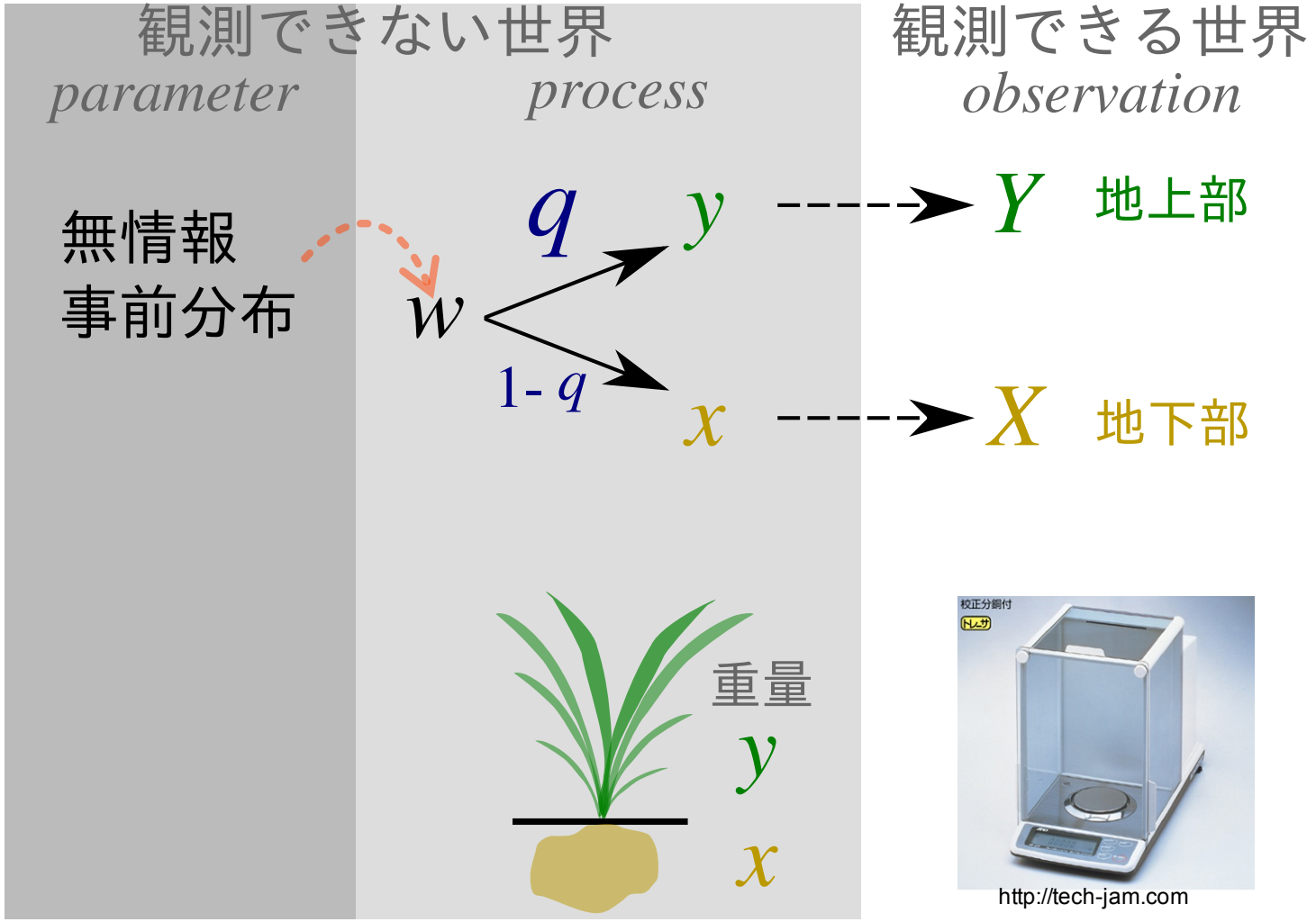
対数スケイルでみるとこうなってます





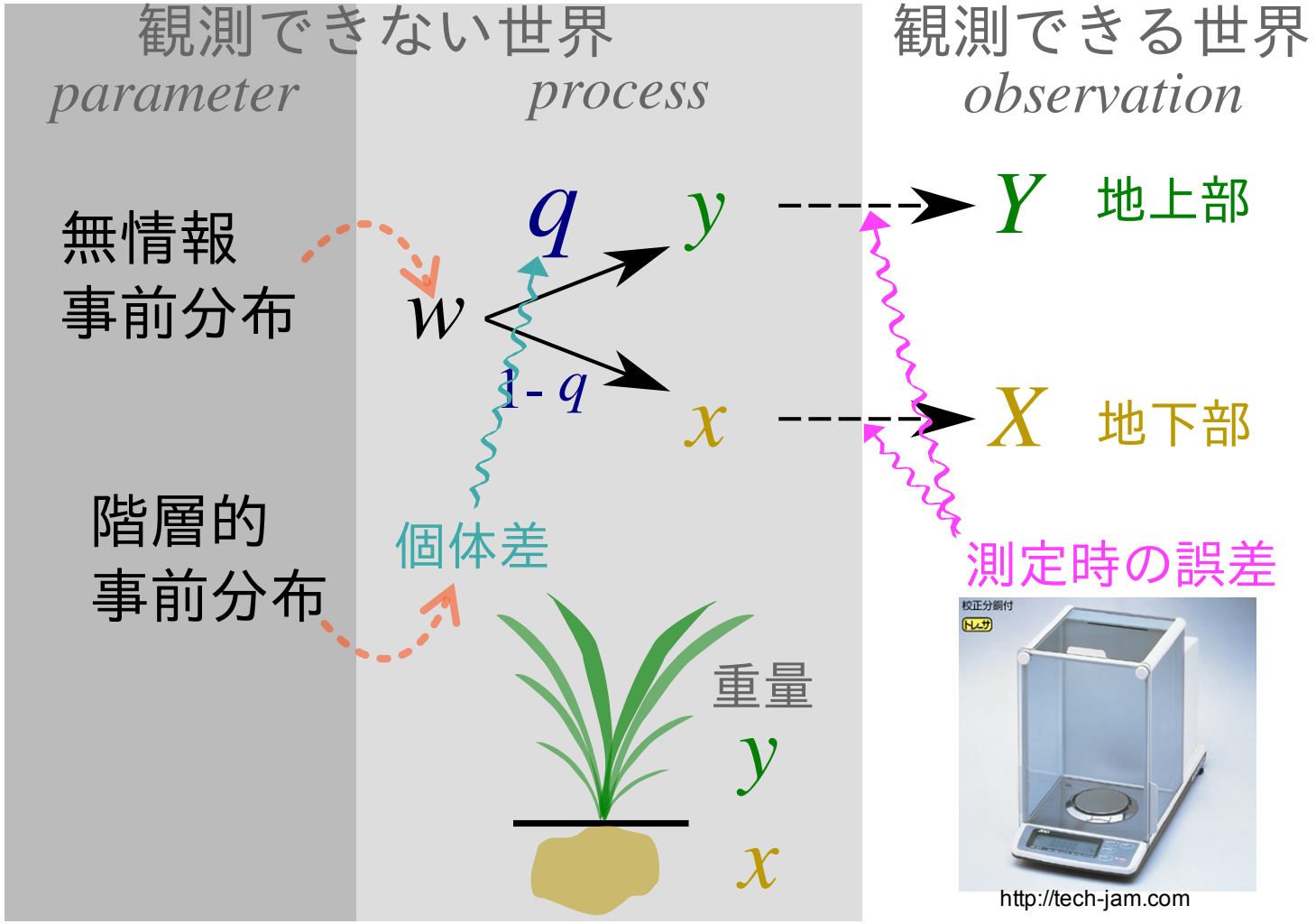
- とはいえ対数世界で「センをひっぱる」わけではない

重量分割モデル (階層ベイズモデル): そのプロセス



http://tech-jam.com

重量分割モデル (階層ベイズモデル): 「誤差」の入りかた



重量分配モデルを BUGS code で (process の部分のみ)

```
for (i in 1:N) {
  Y[i] ~ dnorm(y[i], Tau.err) # 地上部の重量
  X[i] ~ dnorm(x[i], Tau.err) # 地下部の重量
  y[i] <- q[i] * w[i]
  x[i] <- (1 - q[i]) * w[i]
  logit(q[i]) <- a + re[i]
  w[i] <- exp(log.w[i])
  log.w[i] ~ dnorm(0, Tau.noninformative) # !!
}# log.w[i] は地上部 + 地下部の重量
```

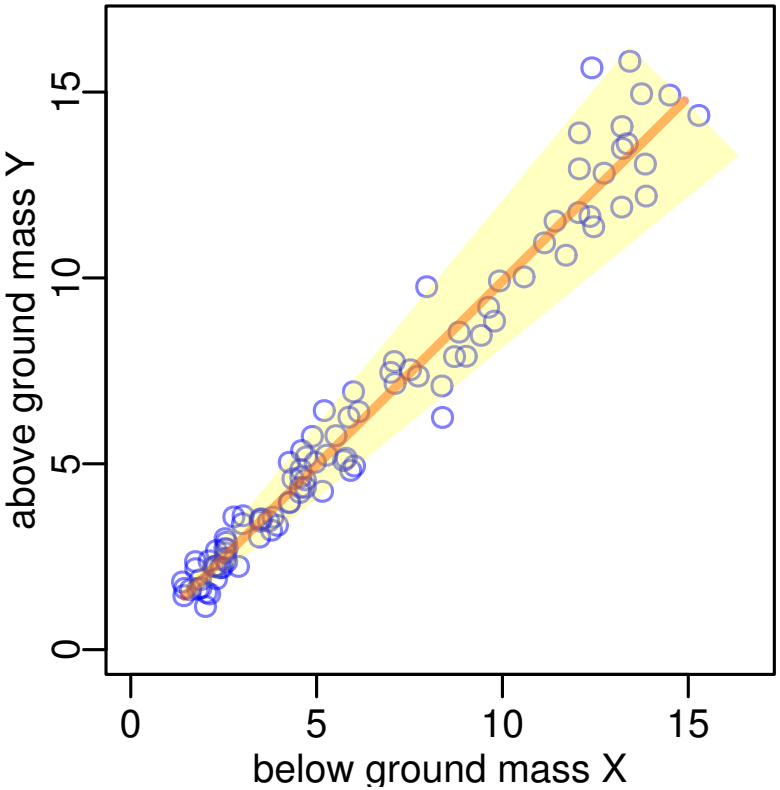
このように明示的にモデルを記述できる!

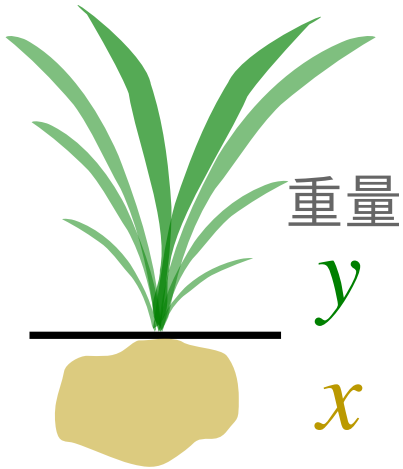
階層ベイズモデルのパラメーター推定: MCMC

1. BUGS code で重量分割モデルを記述する (`model1.txt`)
2. これにデータを渡したりする R スクリプトを書く (`runbus1.R`)
3. R で `runbus1.R` を実行 (`source("runbugs1.R")`)
4. R 内から `library(R2WinBUGS)` によって WinBUGS が起動
5. WinBUGS 内で Markov chain Monte Carlo (MCMC) サンプリング
6. 事後分布からのサンプリング結果が R に渡される

必要なファイルは自由集会サイトからダウンロードできます

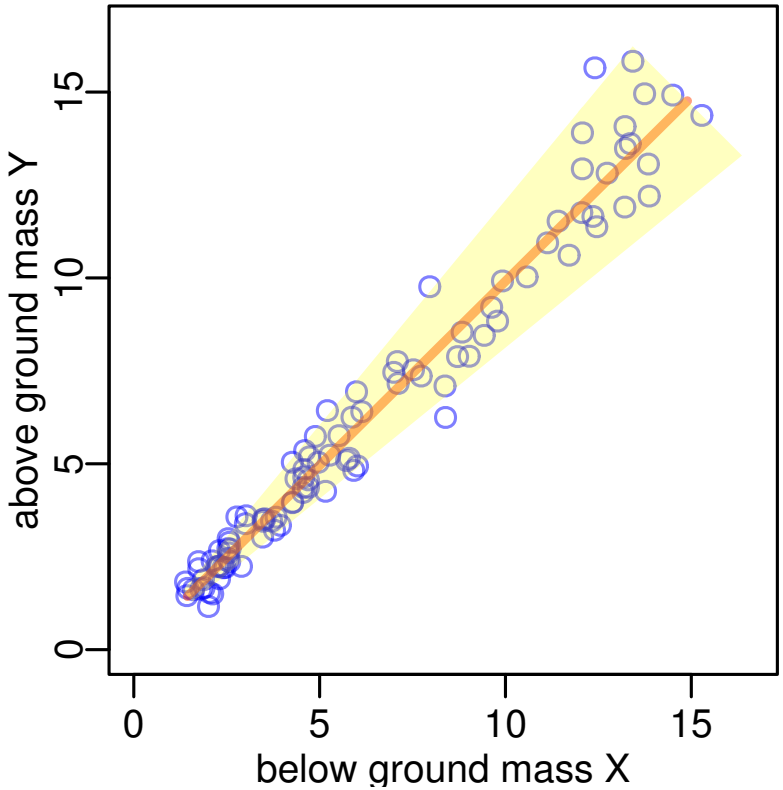
推定結果を組みあわせた予測

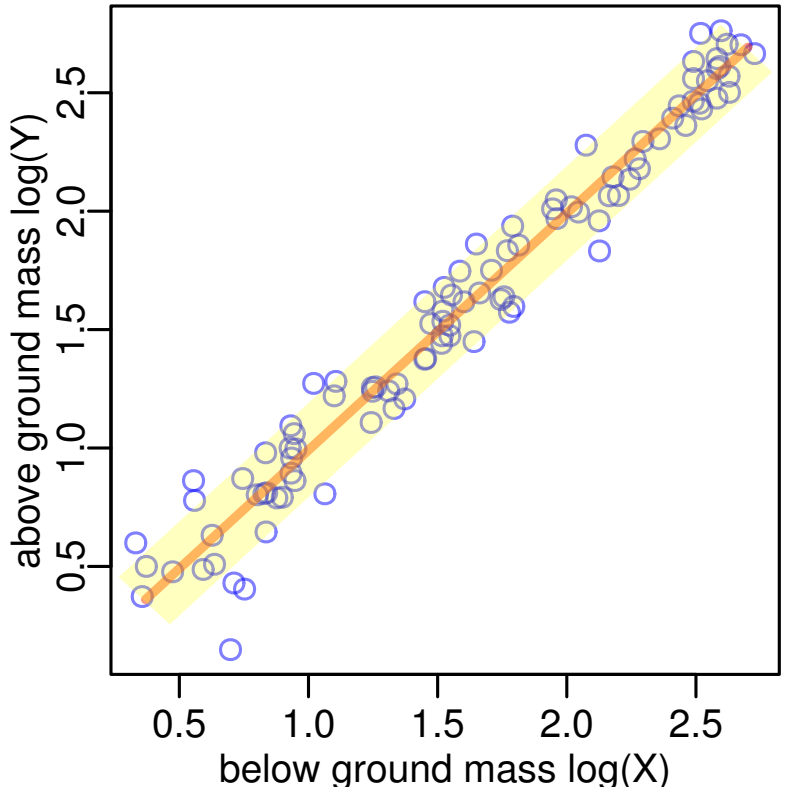




- オレンジ色の線は中央値 (median)
- 黄色の領域は個体差による予測のばらつき (95% CI)

個体差によるばらつき，そして測定時の誤差





- 総重量が小さいときには**測定時の誤差**が相対的に大きく
- 総重量が大きくなると**個体差**が占める割合が大きくなる

このモデルの問題点

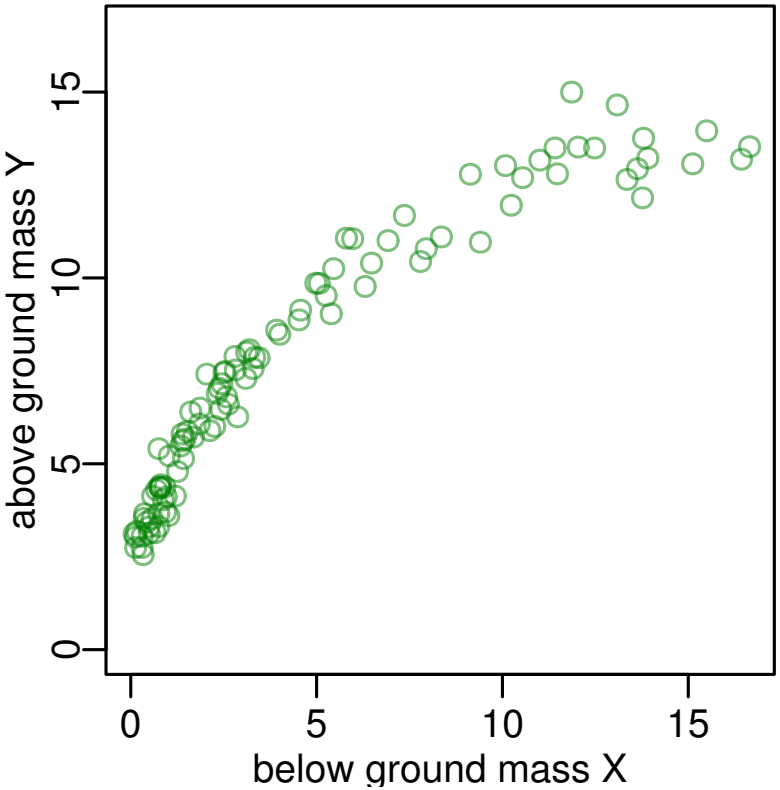
- 測定時の誤差は**正規分布**と仮定していいのか?
  - そうですね，重量は非負の値なのに……
- **測定時の誤差の大きさ**はどうやって推定したの?
  - 今回は「真の値」をほうりこみました
  - 実際には，測定機器のカタログとか見ながら「てきとー」に決めるしかないのかも? (主観的な事前分布)
  - ひとつの観測対象に対して，複数の測定値が得られていれば，階層ベイズモデルで測定時の誤差の大きさを推定できます
- 状況がちょっと単純すぎない?
  - それでは次の例題を……

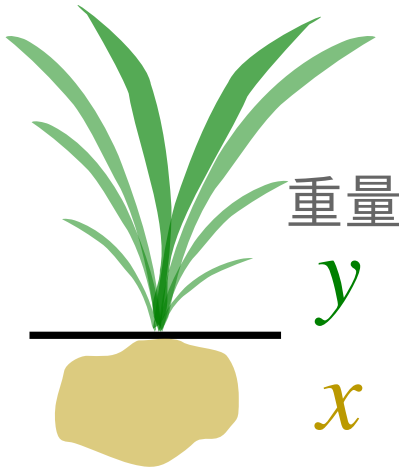
例題 2

もうちょっと複雑な

重量分割モデル

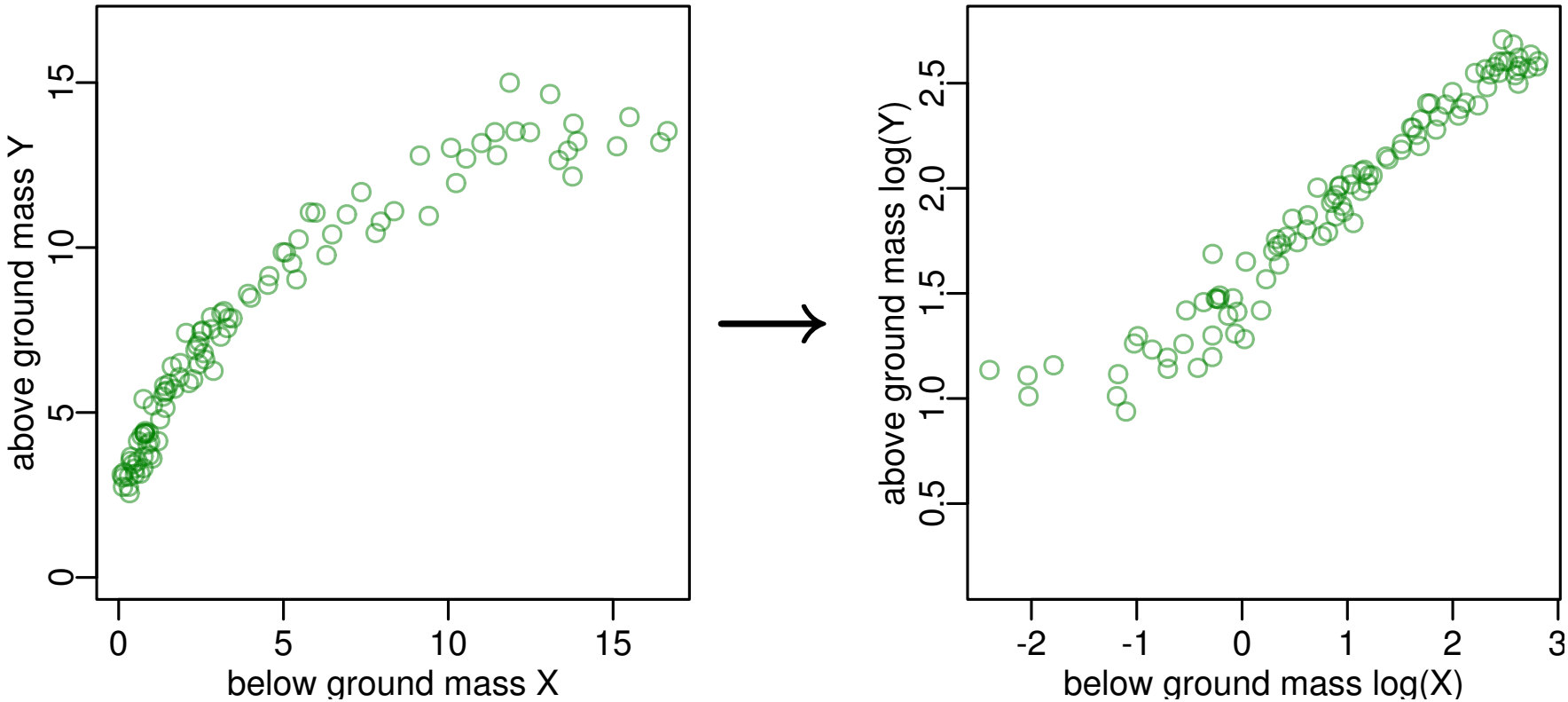
架空データ 2: 重量増大とともに分配が変化





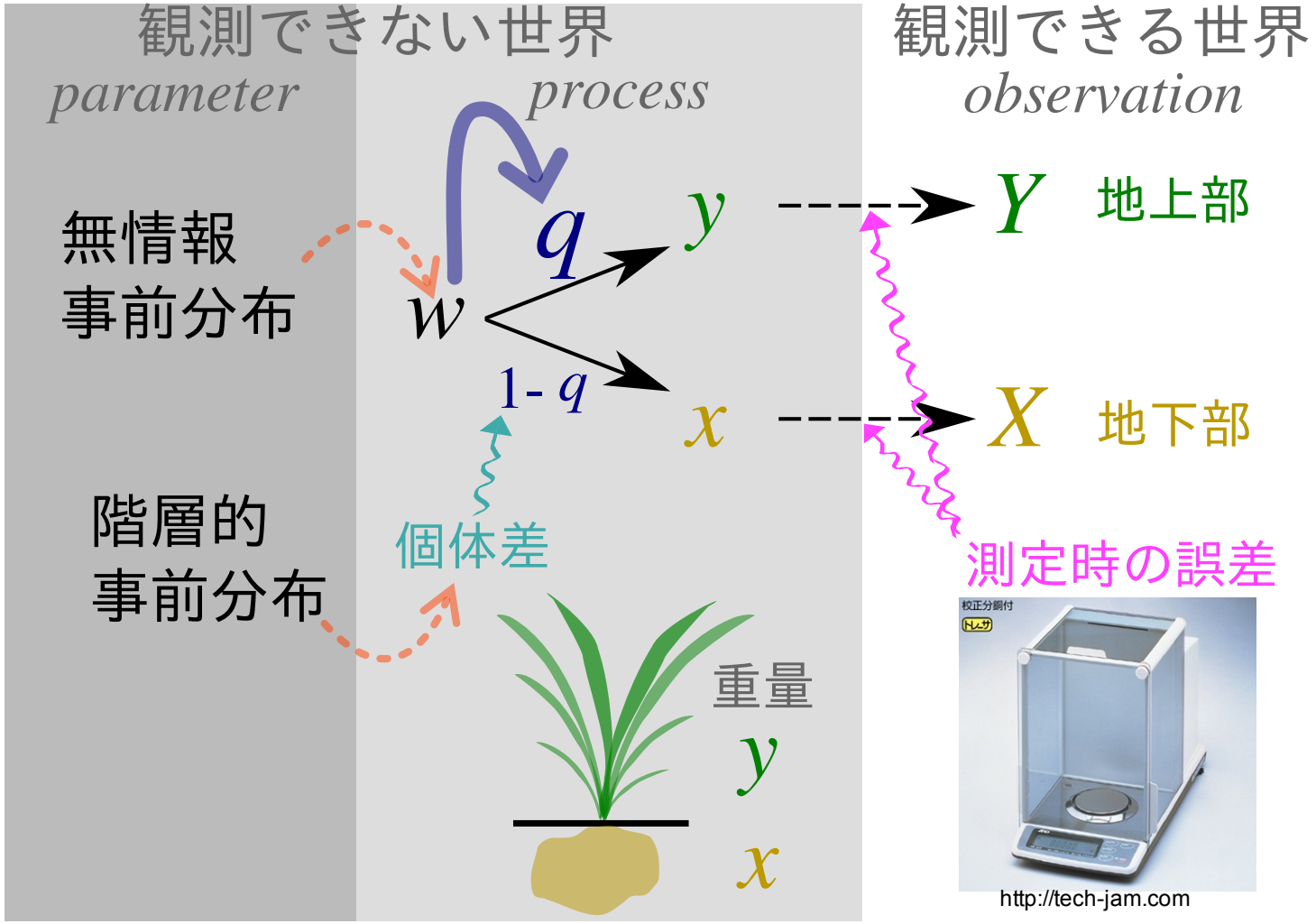
- 小さいときには地上部重量を大きくする
- 総重量が大きくなってくれば地下部を大きくする

これはアロメトリーな問題なのだろうか?



- 両対数で直線になっているのか?
- ま，それはあとで考えることにして……

重量分割モデルの改造: q を w 依存にするだけ



BUGS code の変更点

- 先ほどの簡単な例では (切片) + (個体差) だったが

```
logit(q[i]) <- a + re[i]
```

- ここを以下のように**総重量 (w) 依存**に変更するだけ

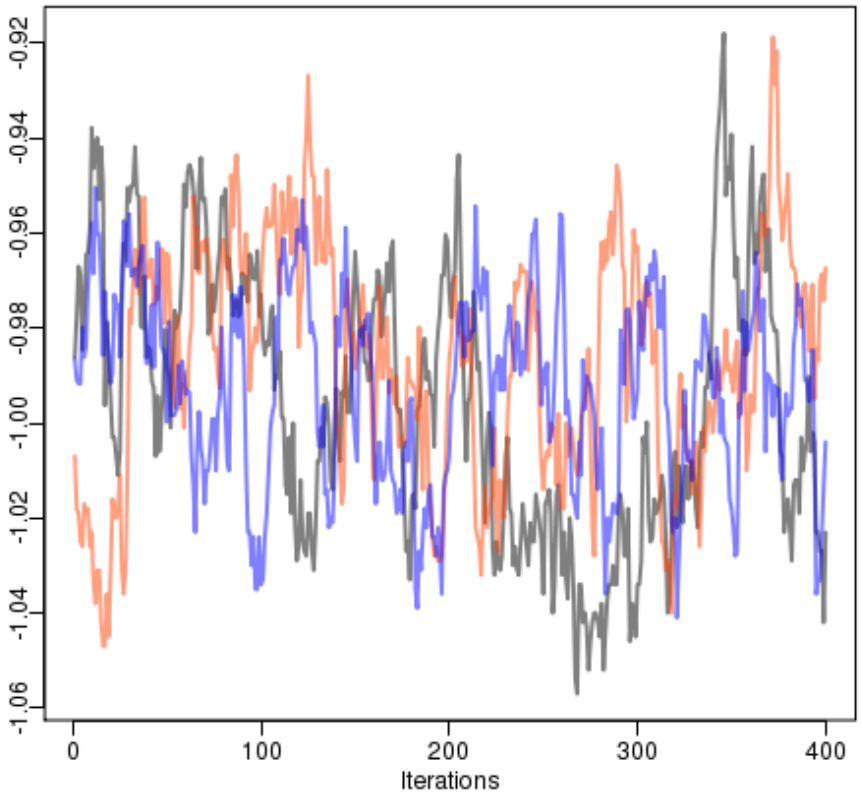
(b ∼ dnorm(0, Tau.noninformative) 追加も必要だけど)

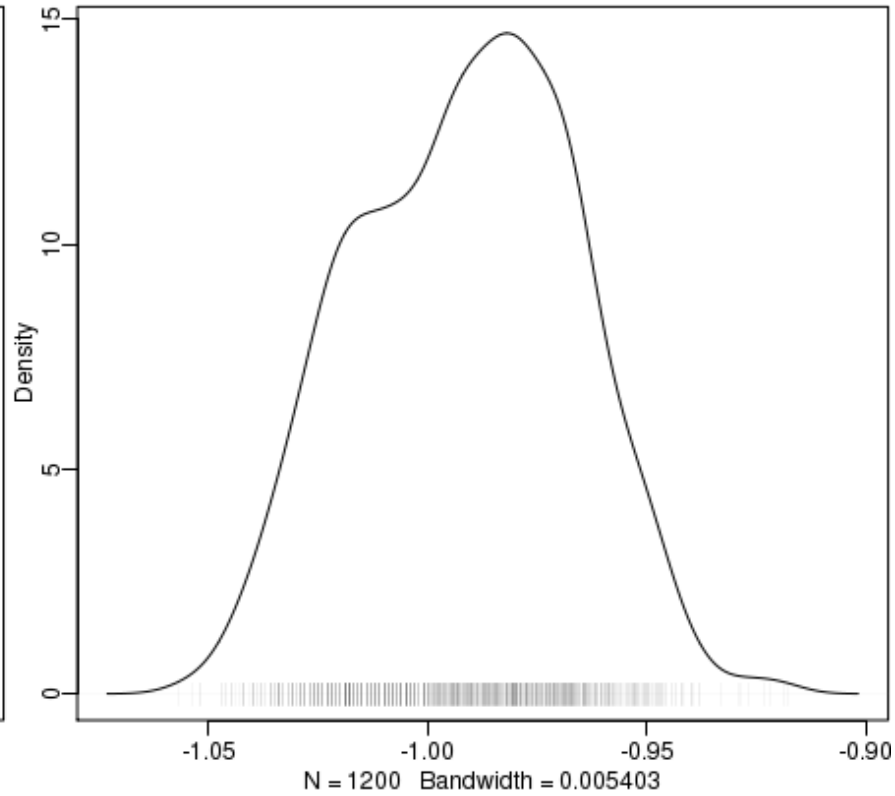
```
logit(q[i]) <- (
   a + b * (log.w[i] - Mean.log.w) + re[i]
)
```

Mean.log.w うんぬんは WinBUGS に必須な中央化ワザ

- あとは R と WinBUGS で MCMC するだけ

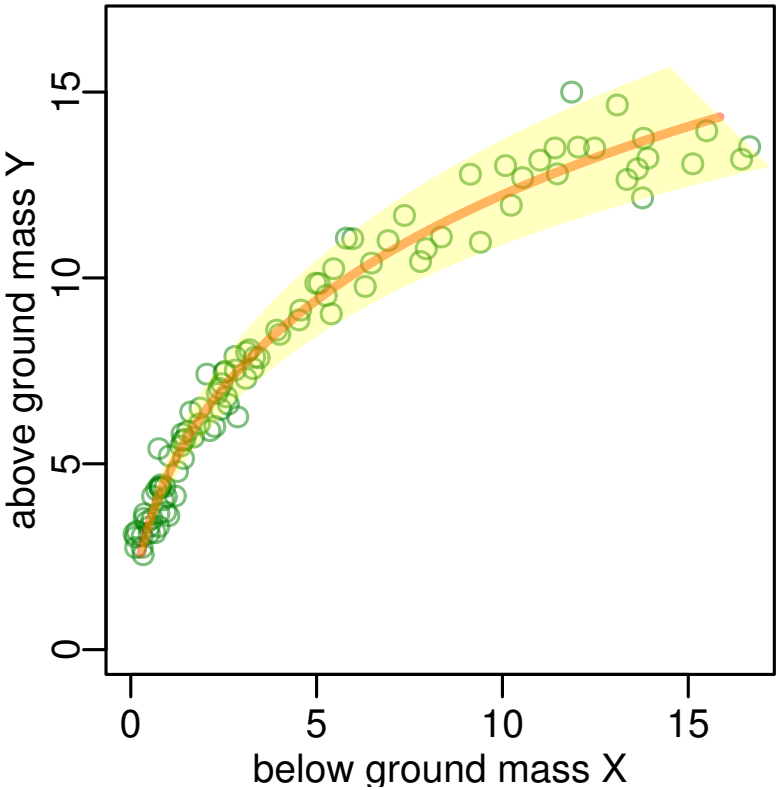
推定結果: 総重量増大 $\longrightarrow$ 地上部への分配減少

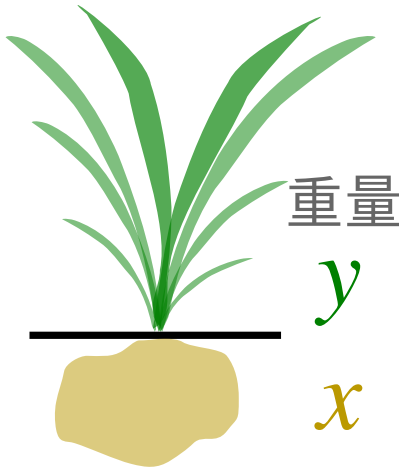




- 総重量 w 依存のパラメーター b はマイナス
- こういう問題は MCMC 収束が遅い

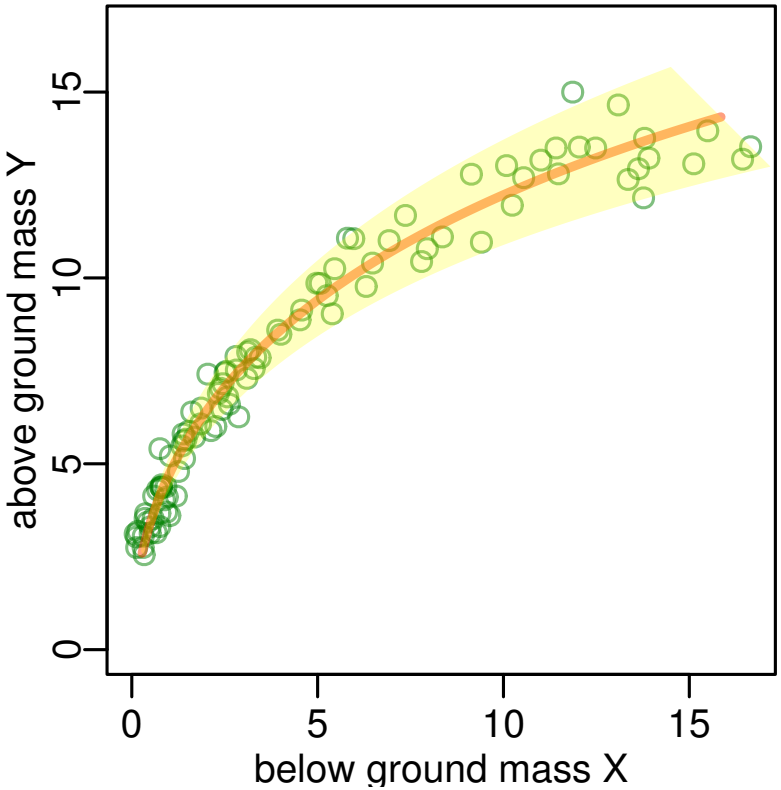
このモデルで複雑な重量分配を表現できる

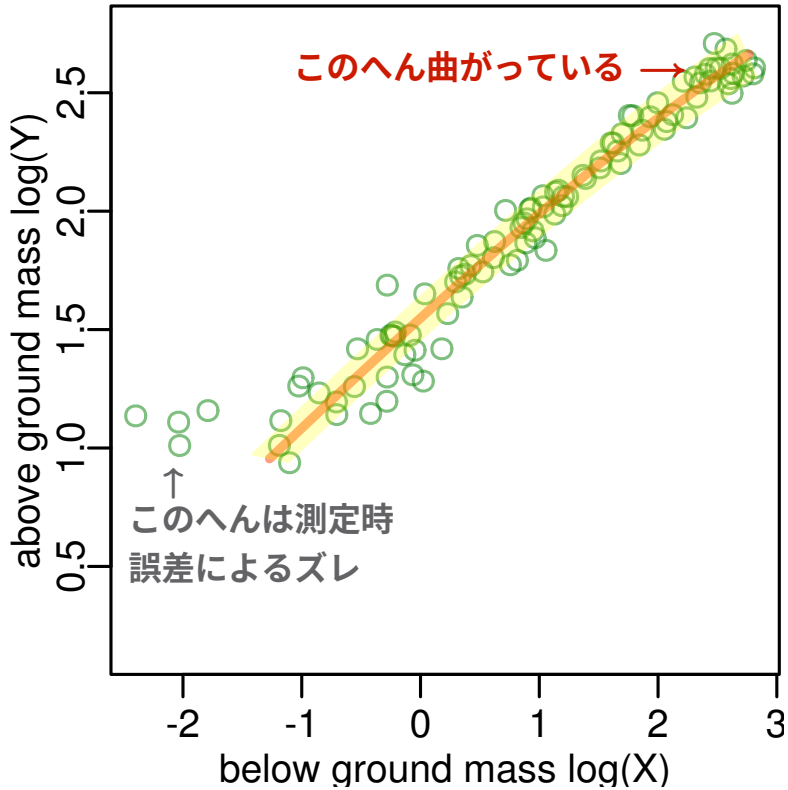




- オレンジ色の線は中央値 (median)
- 黄色の領域は個体差による予測のばらつき (95% CI)

対数の世界でも曲がっている (アロメトリーじゃない!)





- 「両対数表示でも曲がっている」状況でも重量分配モデルは柔軟に対応できる (さらに改訂するのも簡単)

モデルの発展・応用

そしてまとめ

重量分配モデルの発展

- random effects 的な部分の複雑化
  - 個体差だけでなくブロック差・場所差・時系列構造・空間構造……
- 分割数をふたつではなく 3 つ以上にもできる
- 蛇足 1: アロメトリーなモデル (べき乗式) より，重量分配モデルのほうが解釈しやすいモデルではなかろーか?
- 蛇足 2: 応答変数が離散値の場合は二項分布・多項分布モデルで (つまりふつーの二項・多項 logistic 回帰)