

2010-03-15

第 57 回日本生態学会大会 自由集会 (W04)

データ解析で出会う統計的問題 – 「 X の誤差」も統計モデル化

アロメトリーじゃない統計モデル:

X と Y への資源分割

久保拓弥 kubo@ees.hokudai.ac.jp

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

今日の自由集会: ここまでのハナシ

久保による勝手な要約:

- 粕谷さん

- Y だけでなく X にも測定誤差があるときに
- **回帰** をすると係数 (パラメーター) の推定値に偏りが生じる
- その他いろいろ不都合が生じる

- 伊東さん

- ベイズ統計モデリングによって簡単に X の誤差をモデル化できる
- しかしベイズモデル化した**回帰** でも推定の偏りは解消されない
- ひとつの測定対象で複数の測定してみるのが重要だろう

今日の久保バナシ: 先のお二人とはちょっとズレてます

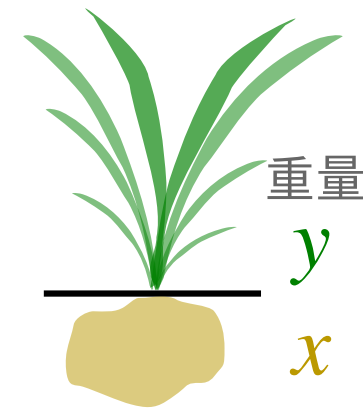
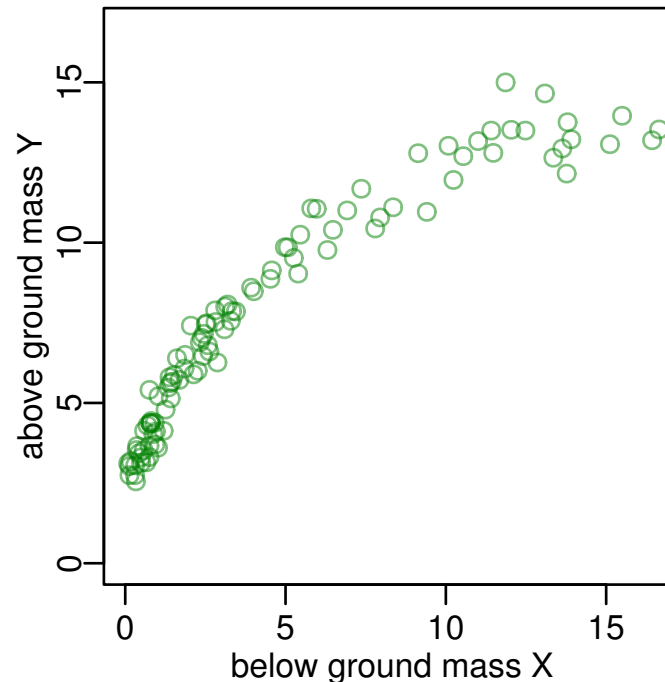
- みんな**回帰**が好きだ
- しかし $\{X, Y\}$ なデータセットを解析するとき, いつも**回帰**でいいんだろうか?
- **回帰**する理由がないのに**回帰**してないか? たとえば, アロメトリーなヒトたち?
- もっと別の**統計モデル**で $\{X, Y\}$ ペアを説明できそう
だ..... たとえば, **重量分割モデル** (久保造語)
- そのモデルの中で $\{X, Y\}$ の**測定誤差**のあつかいも考えてみよう

今日のハナシのながれ

1. まずは $\{X, Y\}$ 誤差ありデータ解析における，ありがちな**回帰適用**お作法の問題点について検討
2. 解決策のひとつになりうる**重量分割モデル**の紹介
 - とても簡単な架空データ例題を使って
3. 重量分割モデルの拡張方法を検討
 - もう少し現実的な架空データ例題を使って
 - モデルの発展や他の問題への応用を考える

たとえばこういうデータがあったとしましょう

架空植物の地下部 (X) と地上部 (Y) の重量



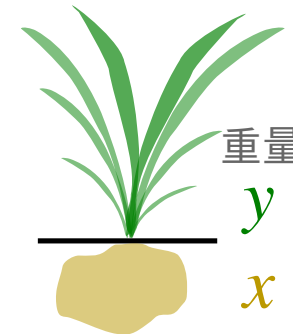
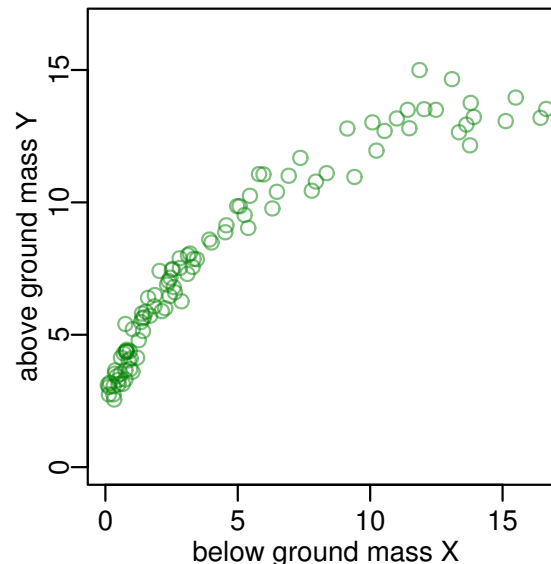
X と Y の関係を調べたい

生態学における $\{X, Y\}$ 誤差ありデータ解析の典型

(今日はずっとこのたぐいの架空植物例題ばかり)

「あろめとらー」たちのお作法!

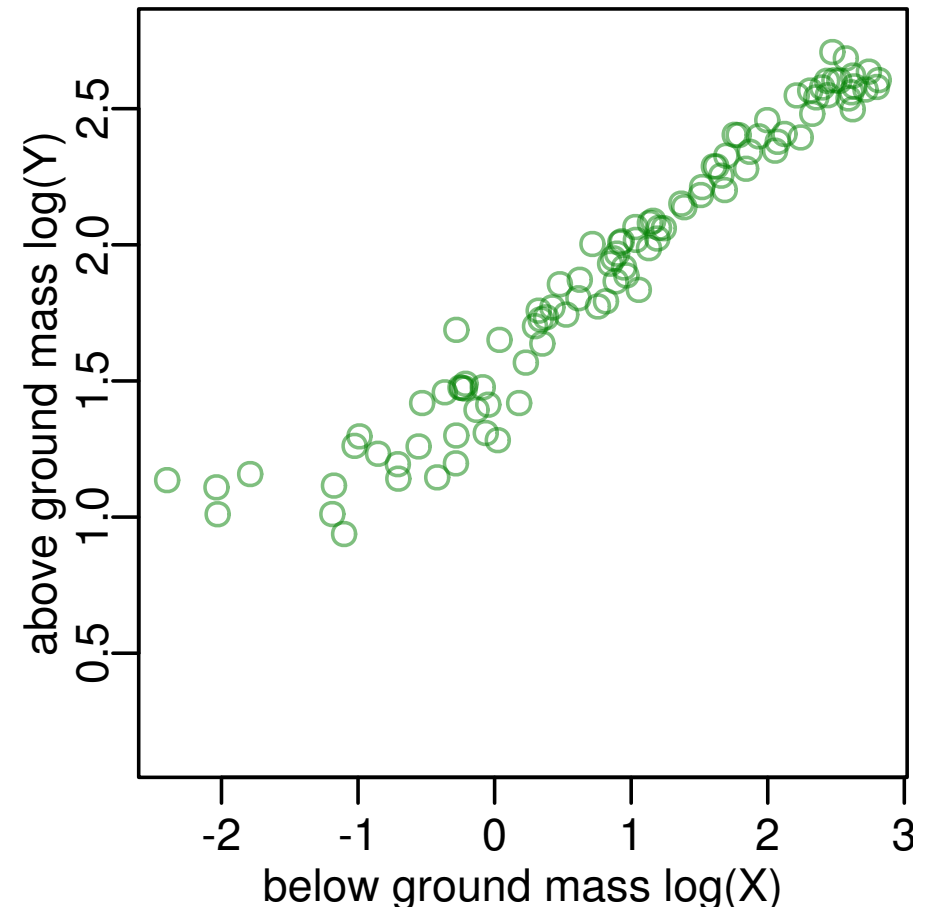
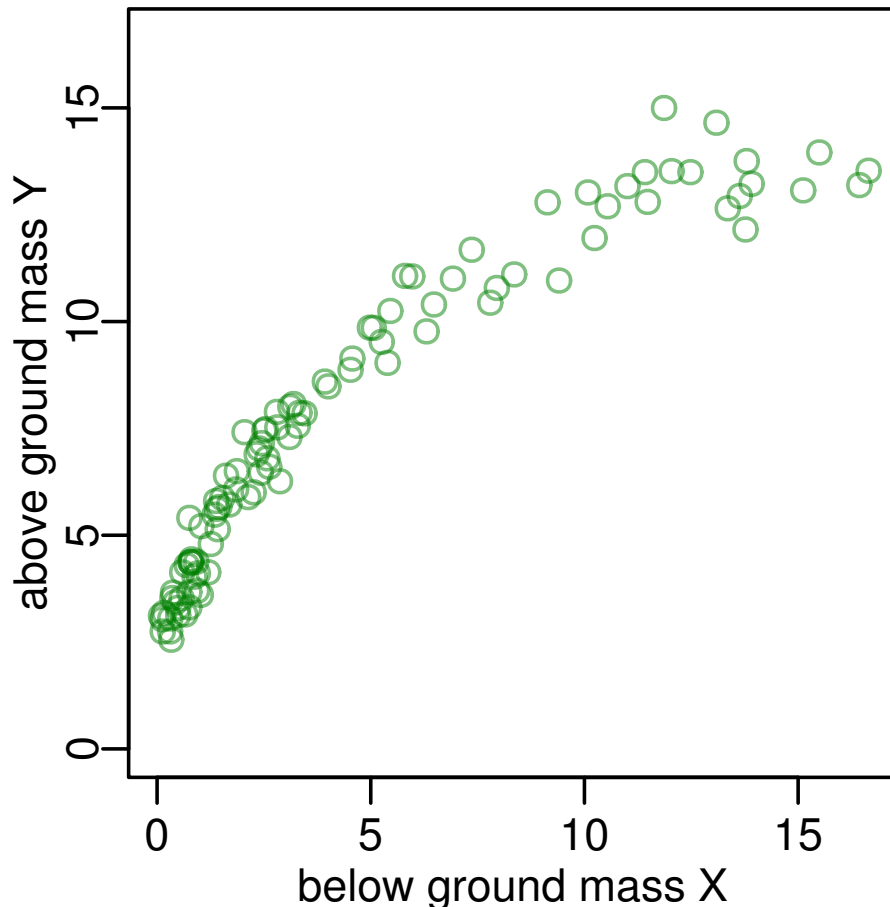
1. X も Y もすぐに対数変換してしまう
2. $\log Y \sim N(a + b \log X, \sigma^2)$ な回帰をやっちゃう
3. 「ゆーい」とか「せつめい力 R^2 」とか言ってみる



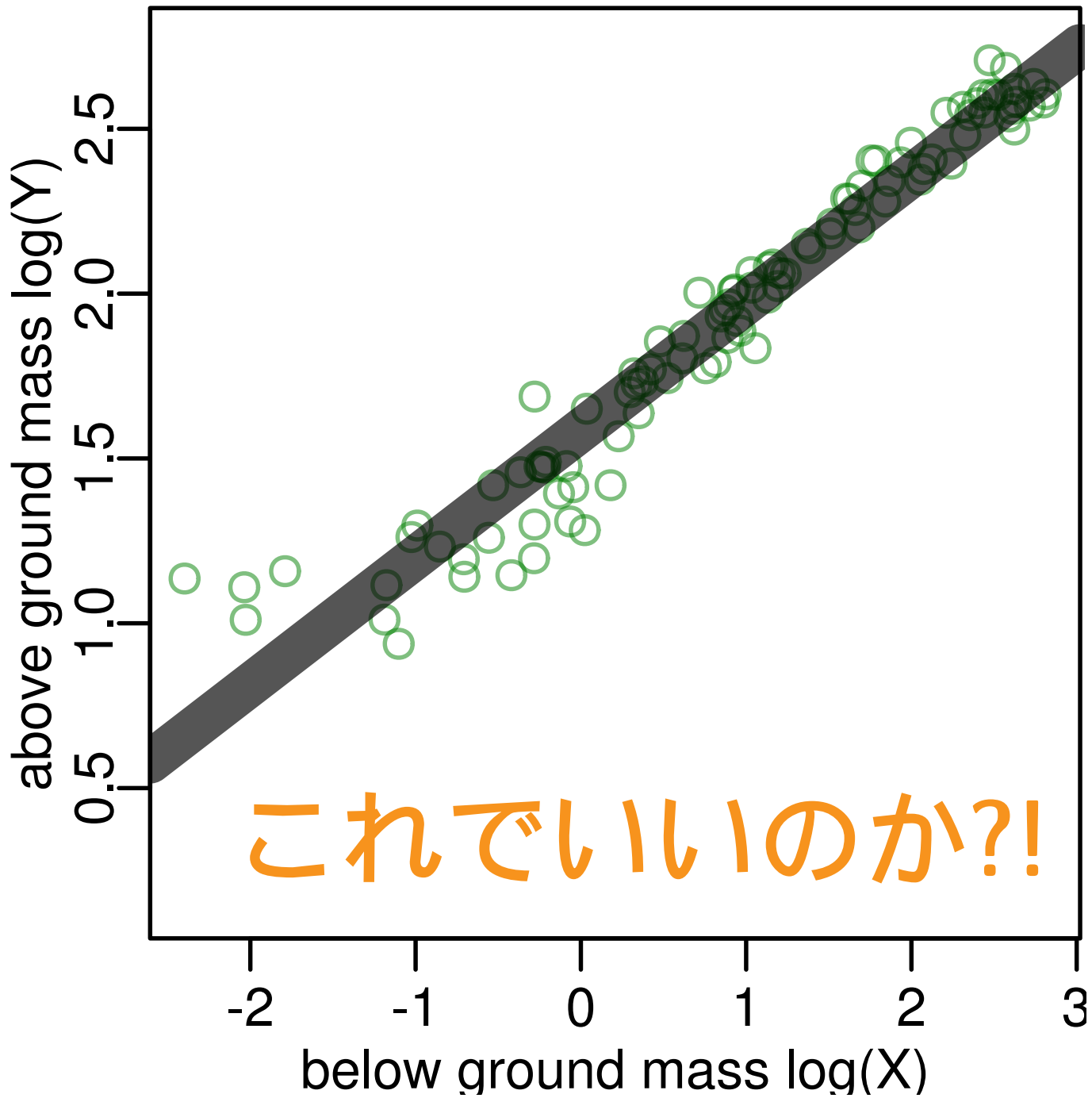
定義

あろめとらー: 上記のようなことをするヒトたち

つまりこのように対数変換してしまつて

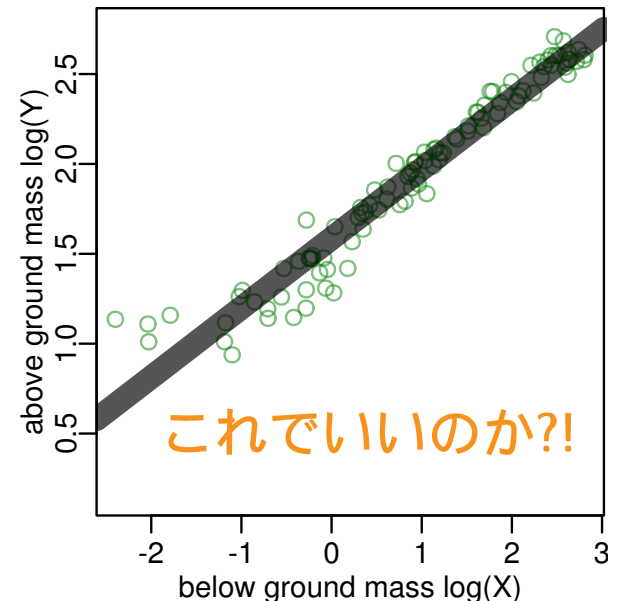


「センをひっぱる」なる行動をします, と.....



あろめとらーなお作法の問題点

- それって地下部重 (X) が**原因**で地上部重 (Y) が**結**果なのか?
 - 因果関係がわからんのに**回帰**してよいの?
- なぜ $E(\log Y) = a + b \log X$?
 - これって何を表現しているモデル?
- X の測定時の**誤差**はどこにいった?
 - なぜ Y 軸方向にだけ「誤差」?



こういう現象をあつかうもうひとつのモデル?

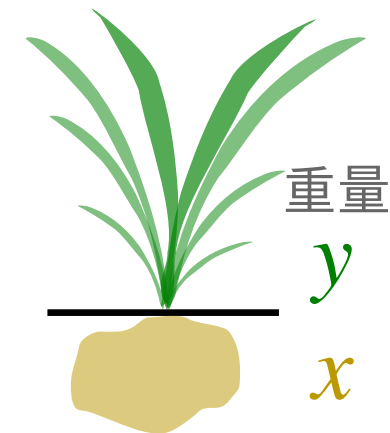
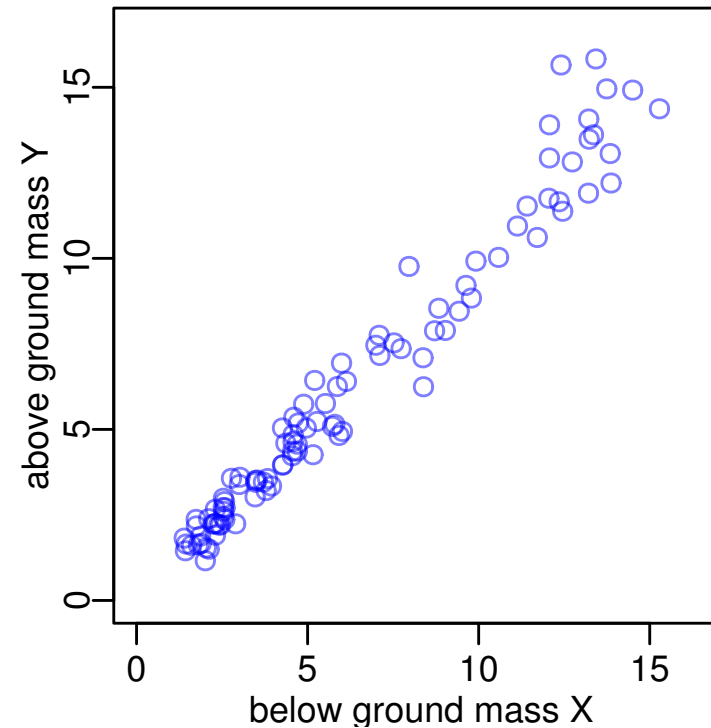
重量分割モデル: 階層ベイズモデルとして定式化

- X も Y も結果である, 原因ではない
 - (バイオマスの) 重量分割という現象の結果
- 全重量 \rightarrow 地下部 (X) + 地上部 (Y) と考える
 - これも「近似的な」現象のとらえかたではあるけれど
- X と Y の測定時の誤差を明示的にあつかう
 - そして「個体差」由来のばらつき (random effects) も考慮する

例題 1

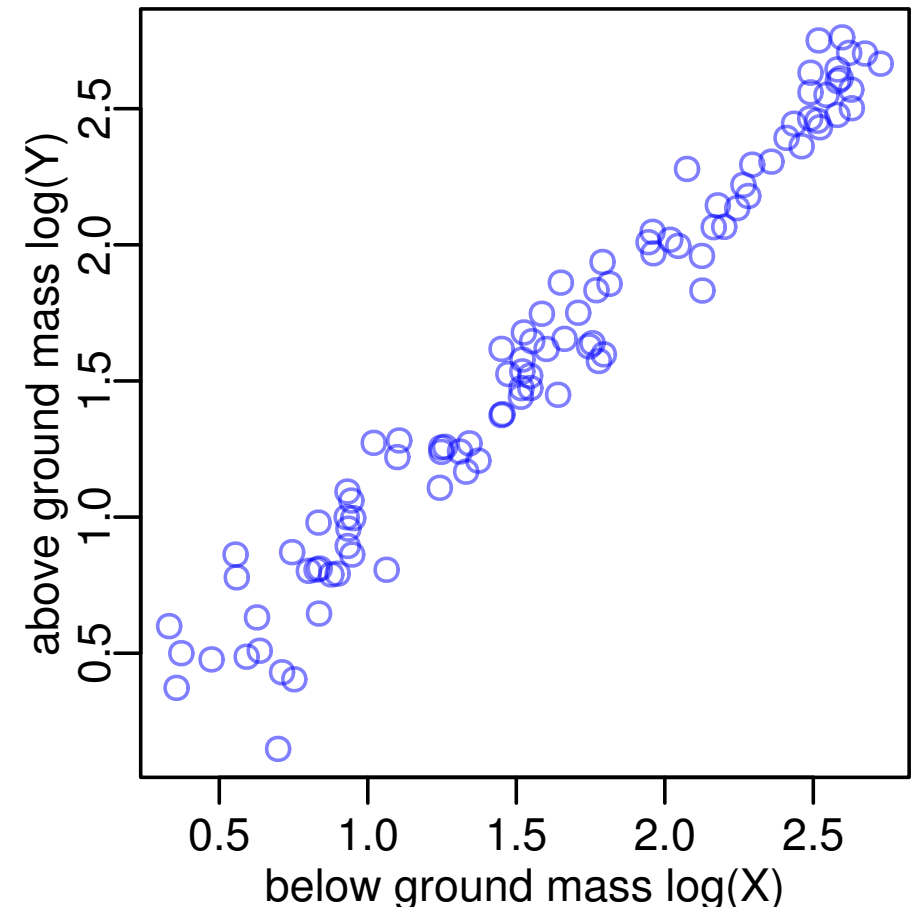
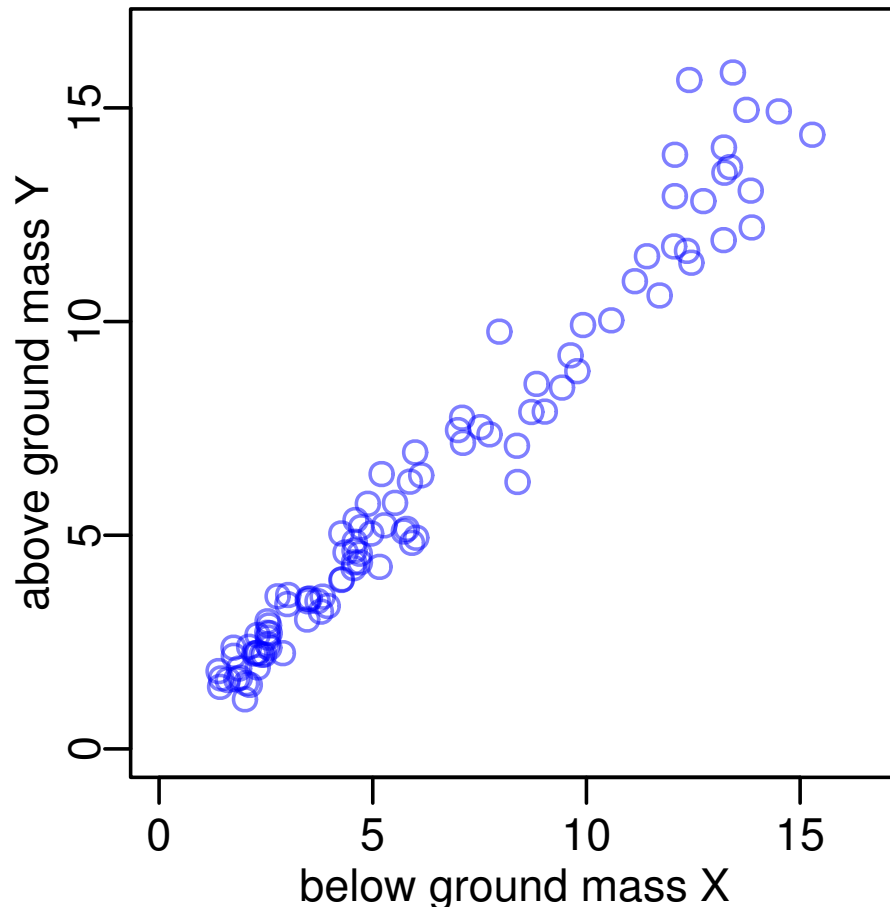
「アロメトリーな回帰」はやめて
重量分割モデルを作ってみよう

架空データ 1: 地下部・地上部の重量



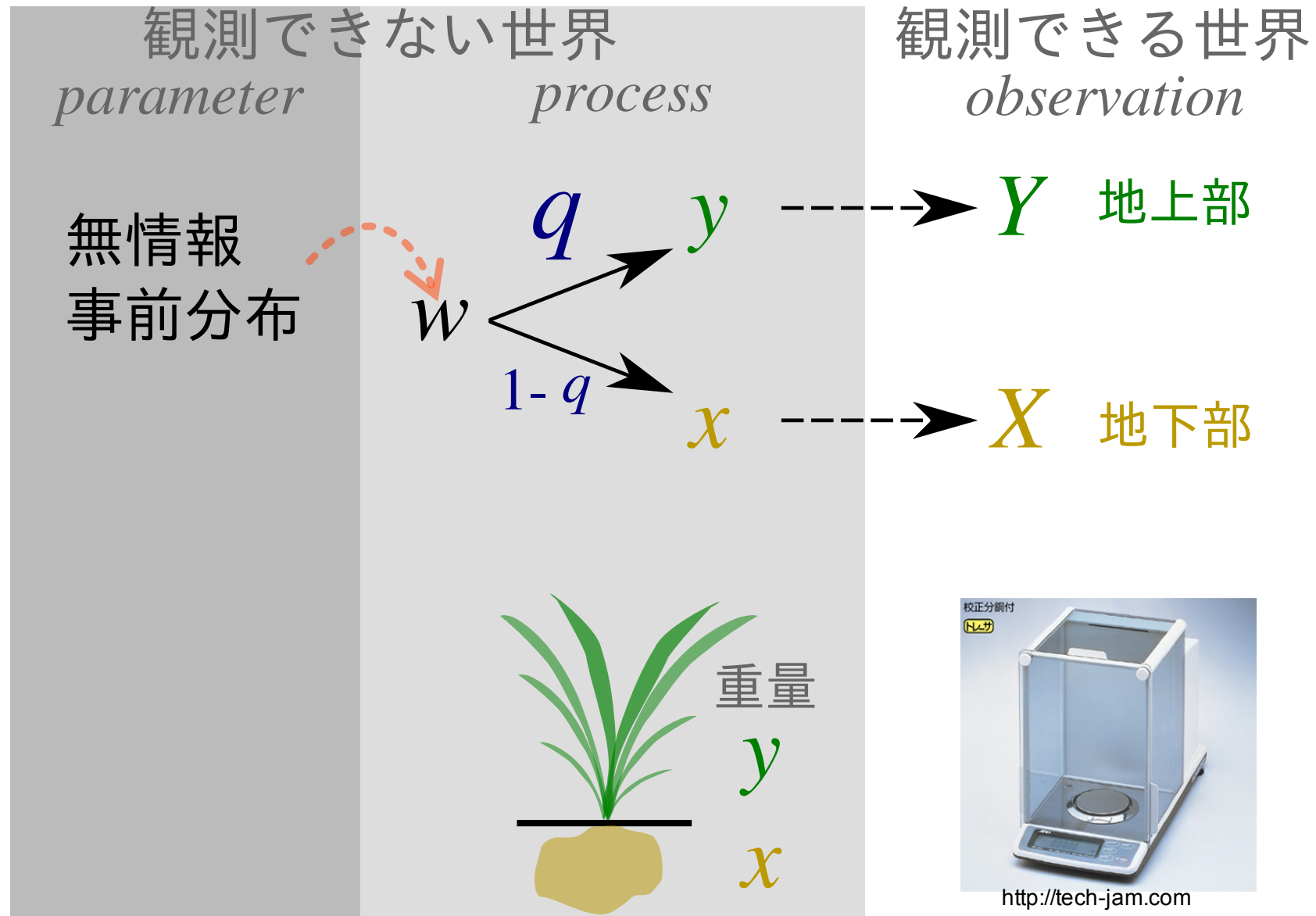
- 対数をとらなくても「直線」にのってる?
- つまり, むしろアイソメトリック (isometric)?
- 重量が重くなるとばらつきが大きくなる?

対数スケールで見るとこうなってます

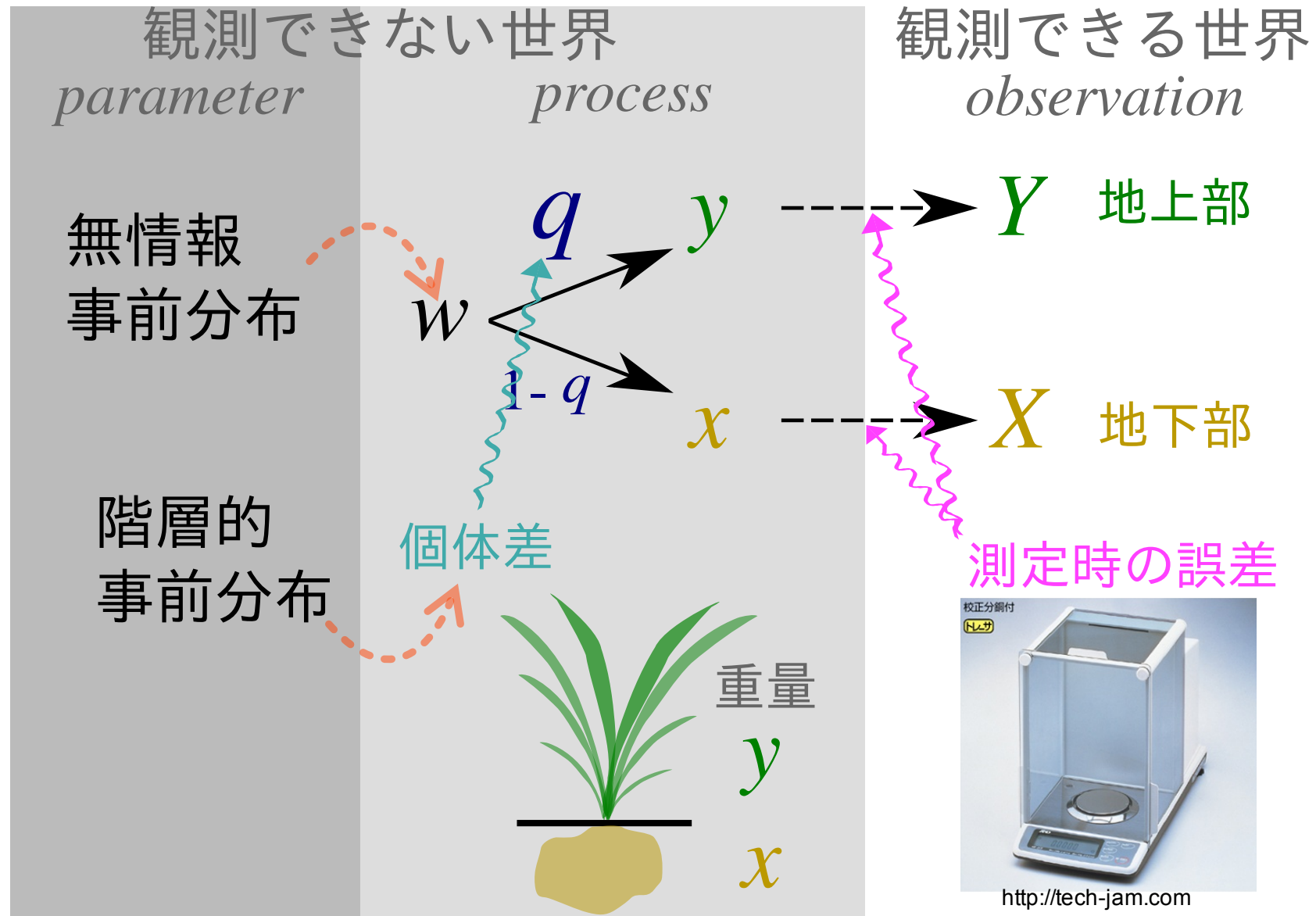


- とはいえ対数世界で「センをひっぱる」わけではない

重量分割モデル (階層ベイズモデル): そのプロセス



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
```

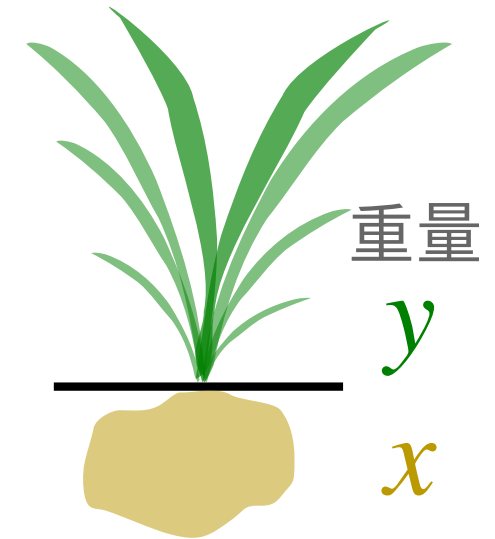
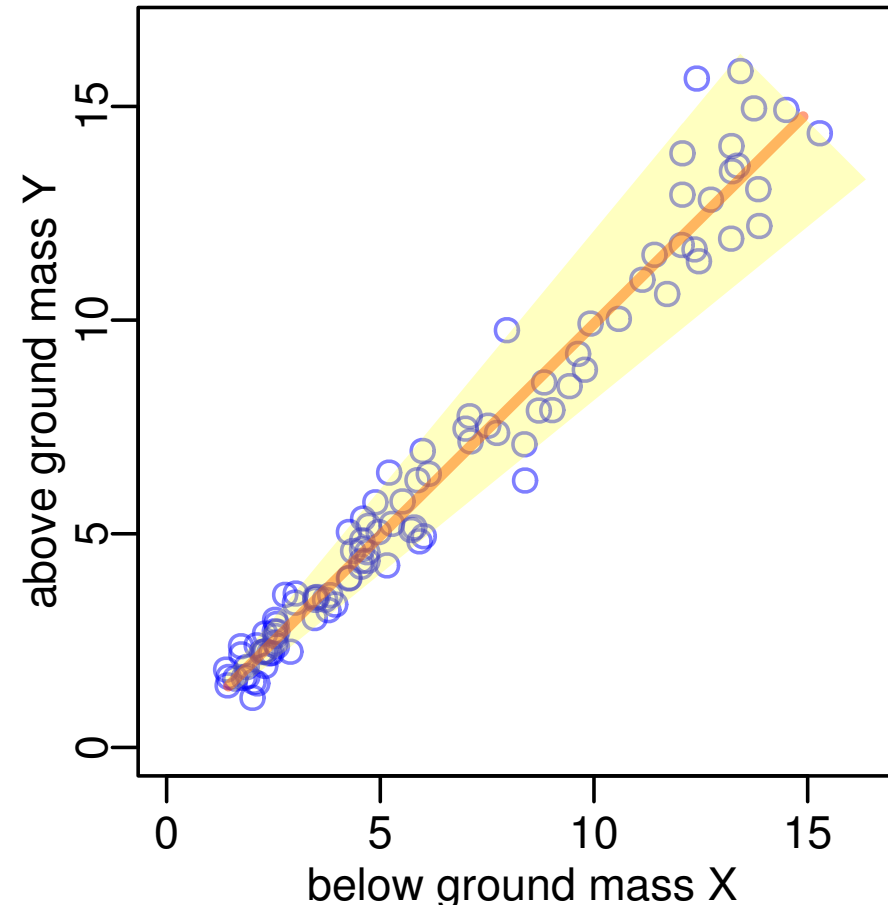
このように明示的にモデルを記述できる!

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) サンプルング
6. 事後分布からのサンプリング結果が R に渡される

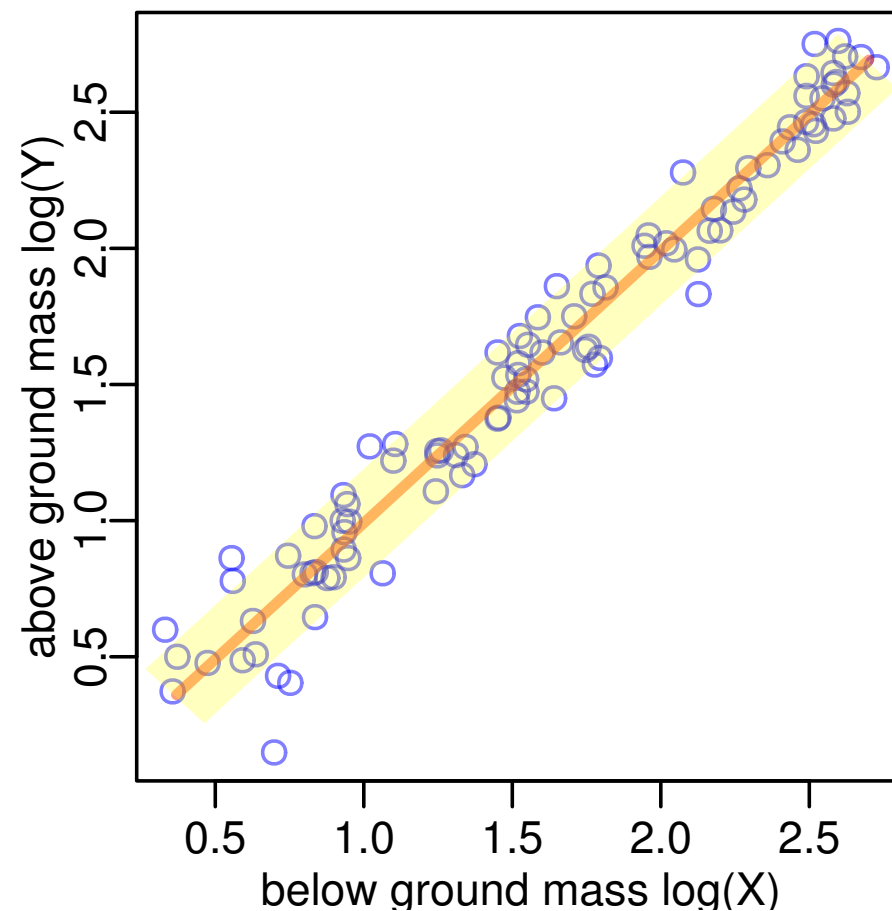
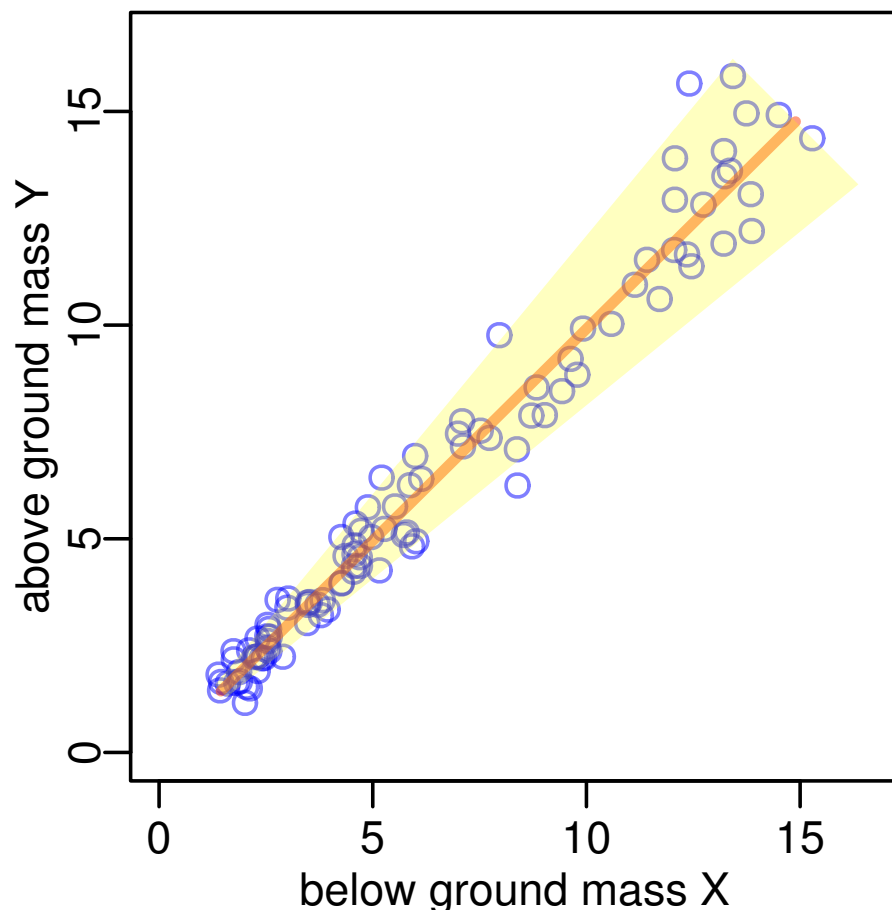
必要なファイルは自由集会サイトからダウンロードできます

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

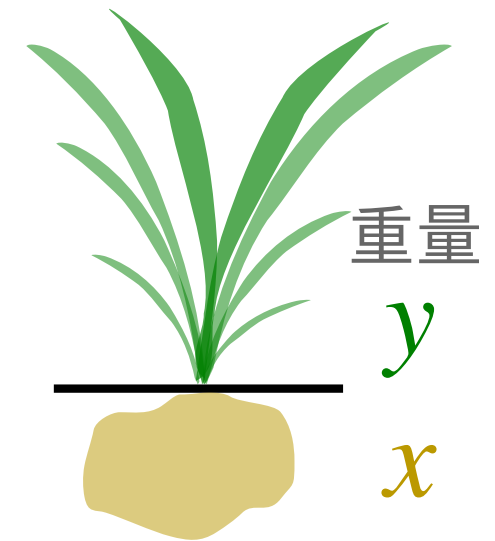
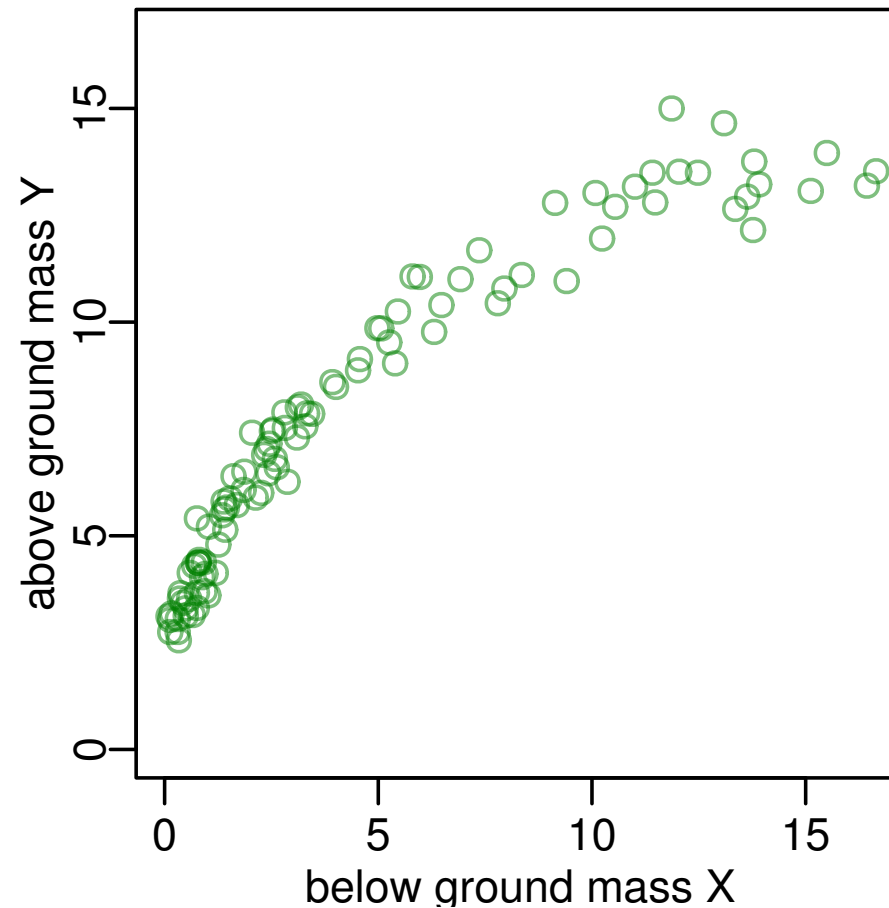
このモデルの問題点

- 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

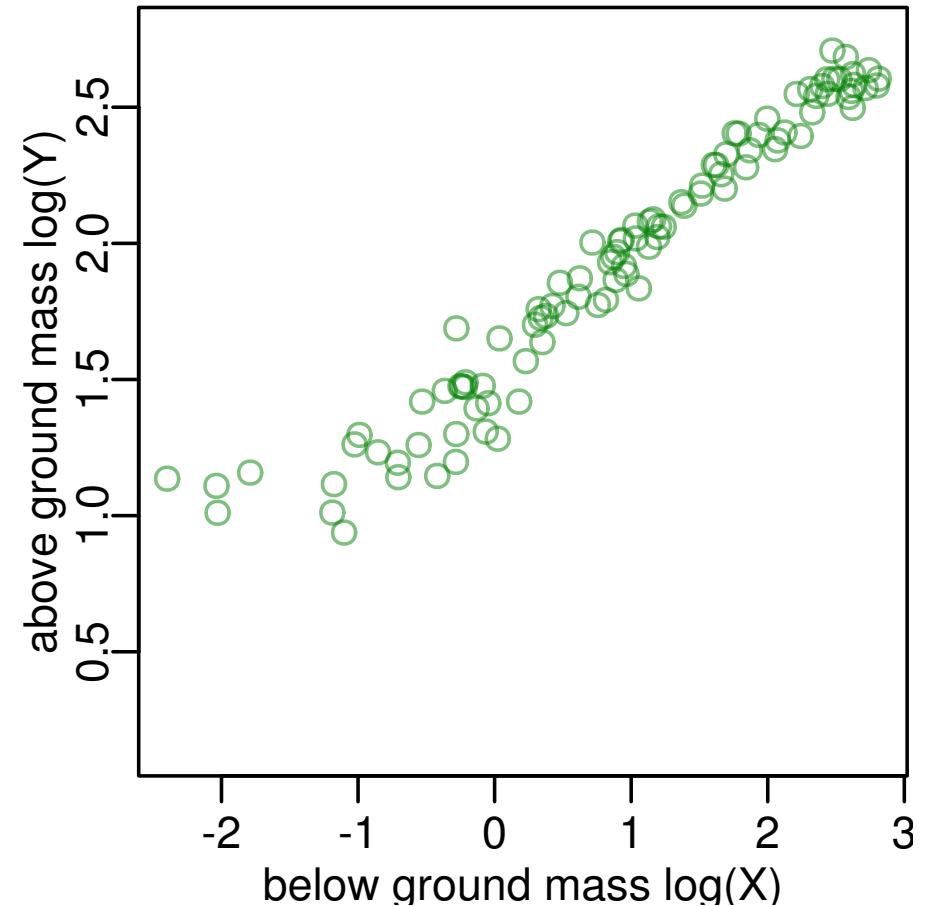
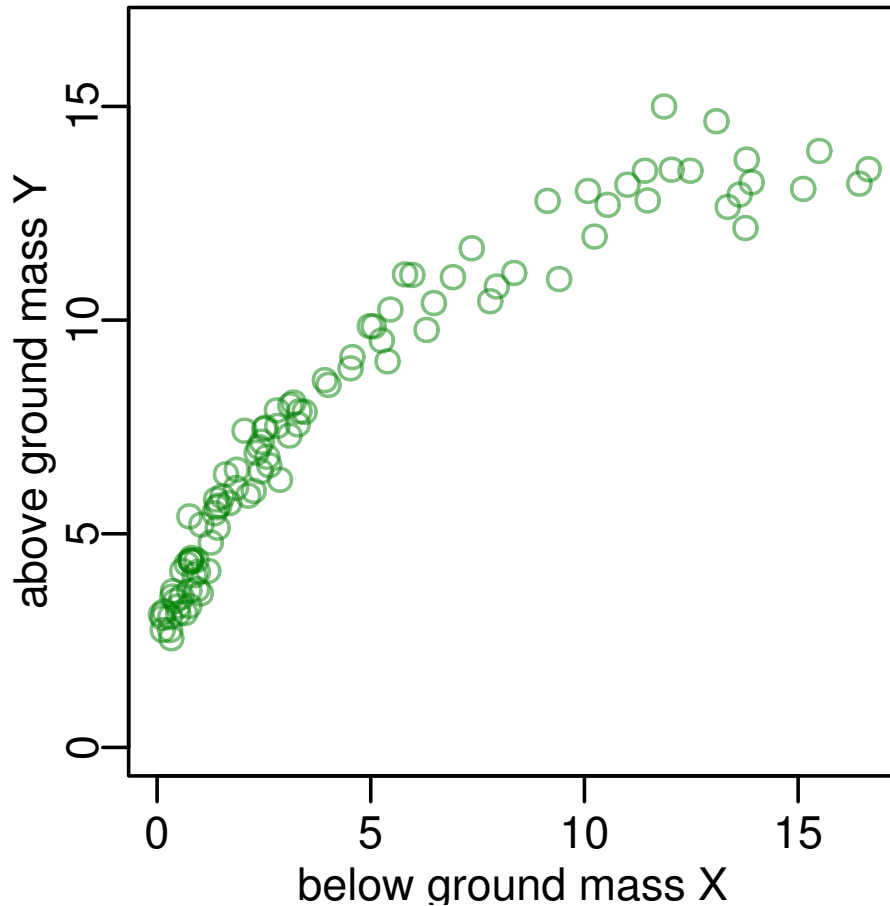
もうちょっと複雑な 重量分割モデル

架空データ 2: 重量増大とともに分配が変化



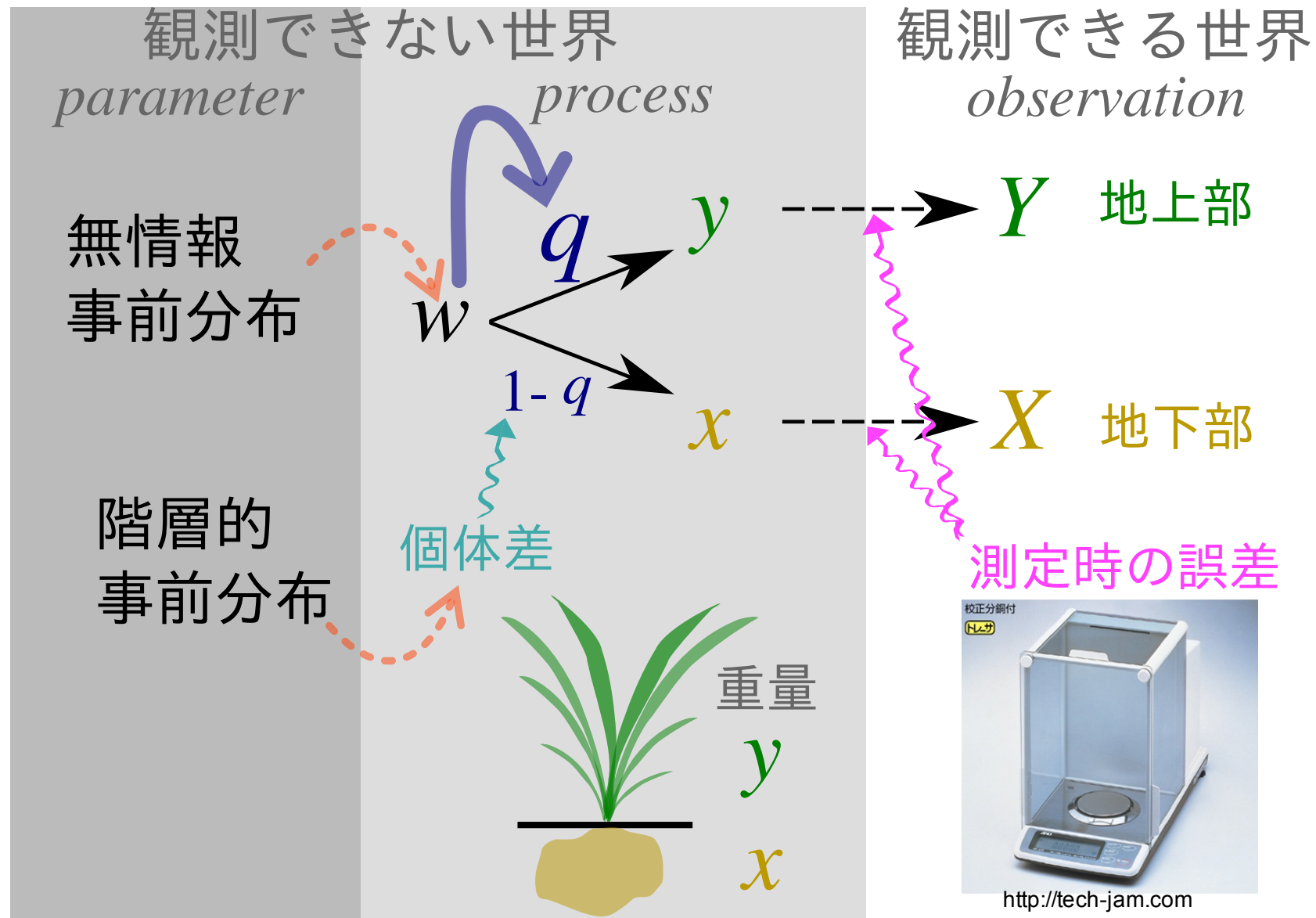
- 小さいときには地上部重量を大きくする
- 総重量が大きくなってくれば地下部を大きくする

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

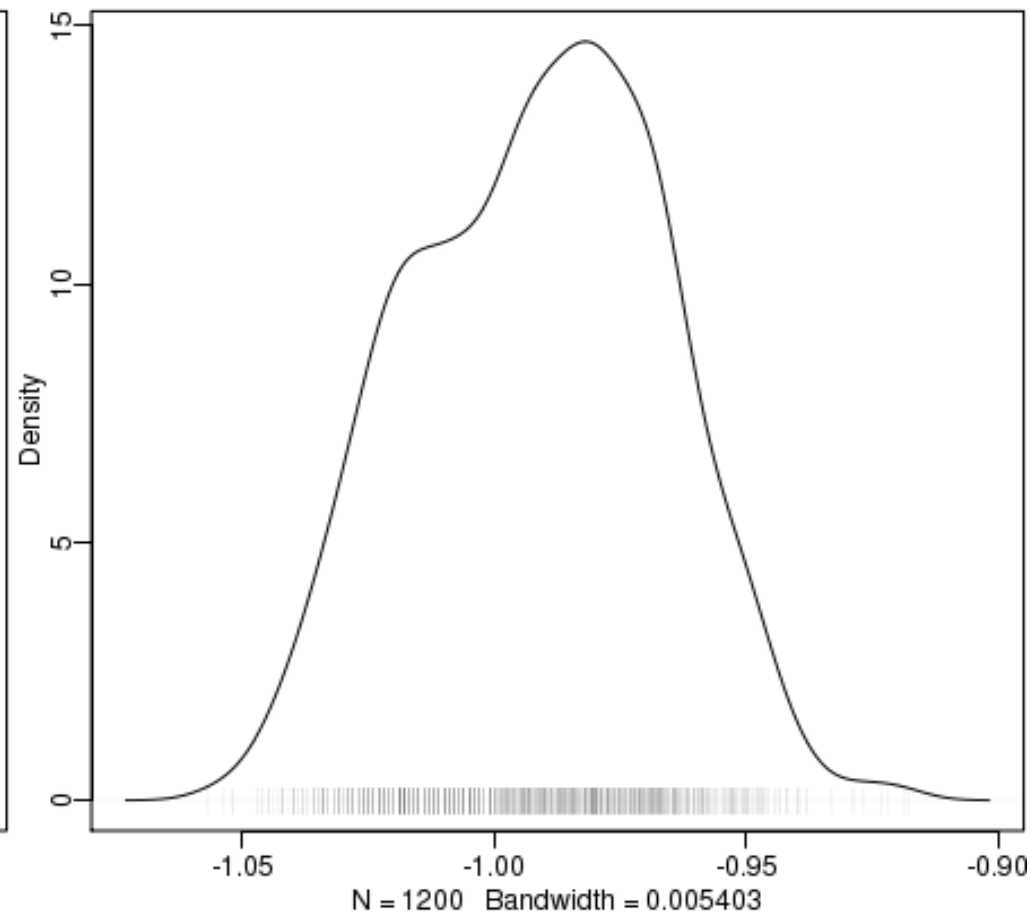
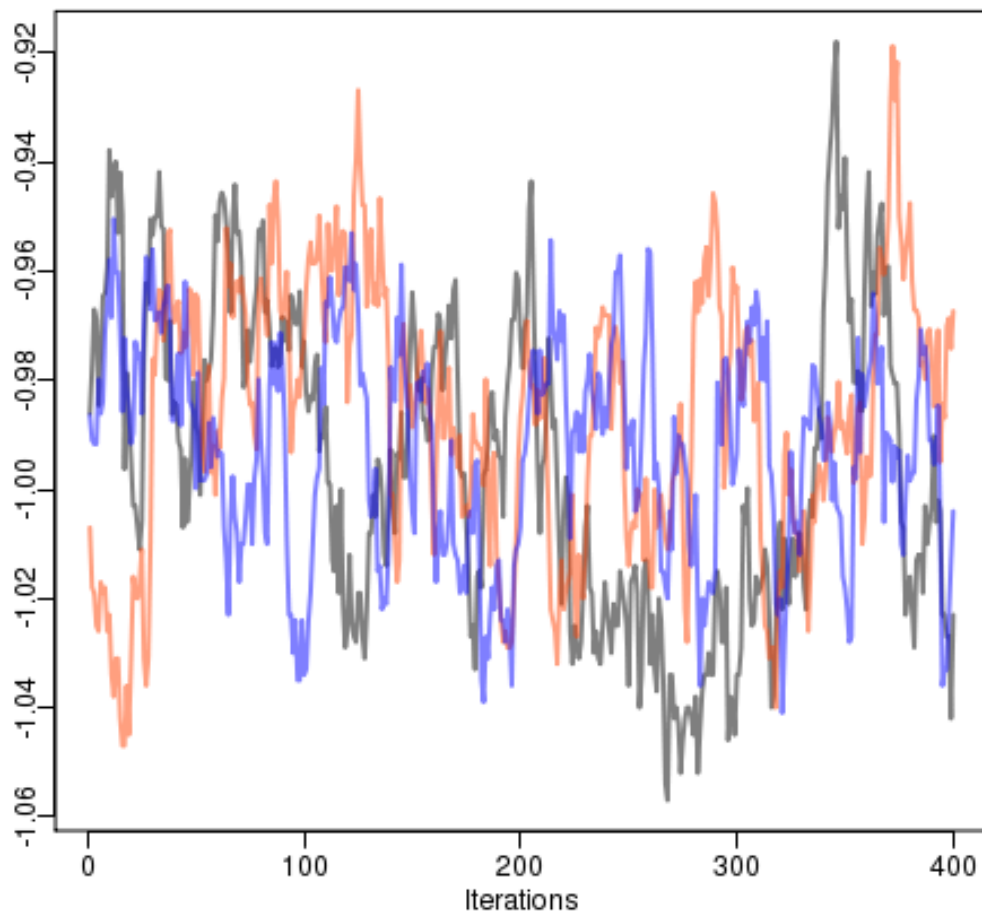
($b \sim \text{dnorm}(0, \text{Tau.noninformative})$ 追加も必要だけど)

```
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

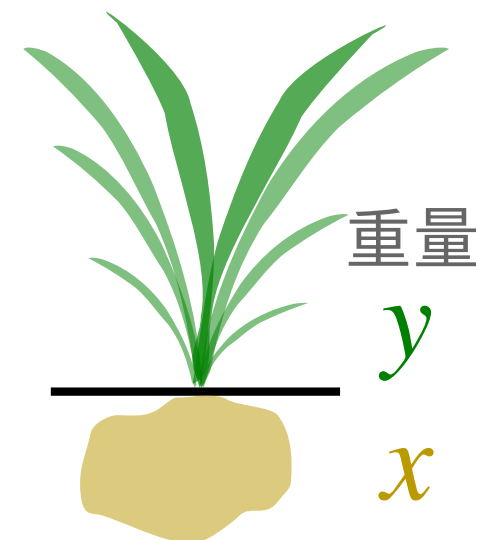
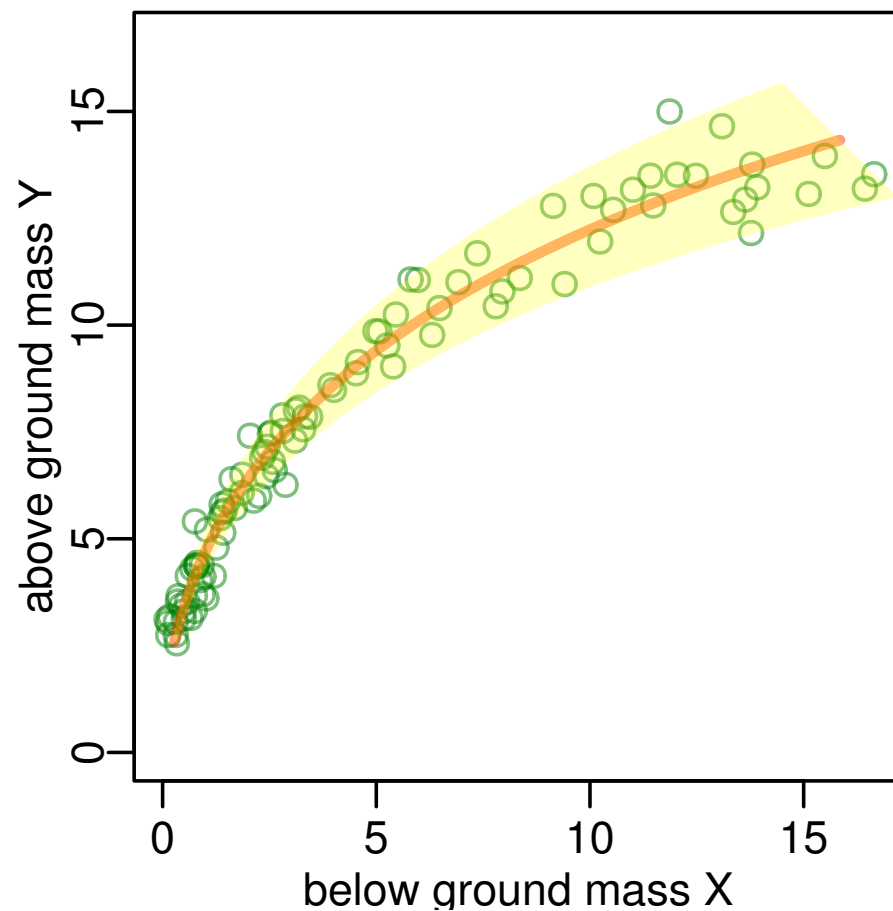
- あとは R と WinBUGS で MCMC するだけ

推定結果: 総重量増大 → 地上部への分配減少



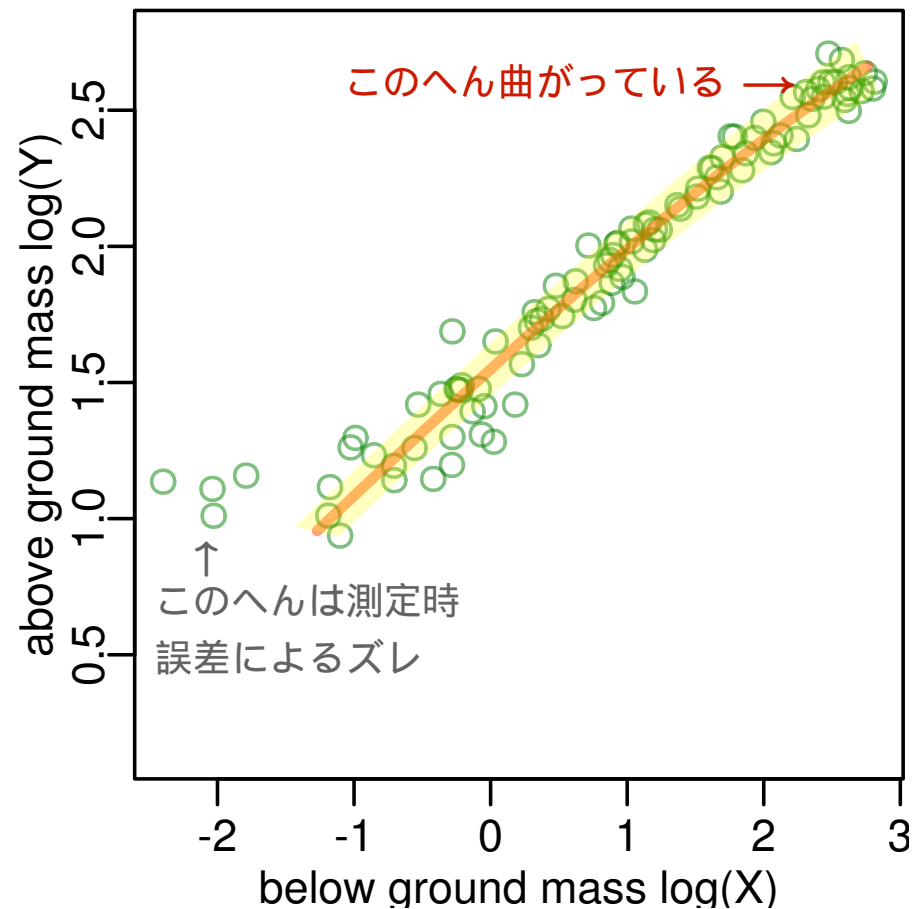
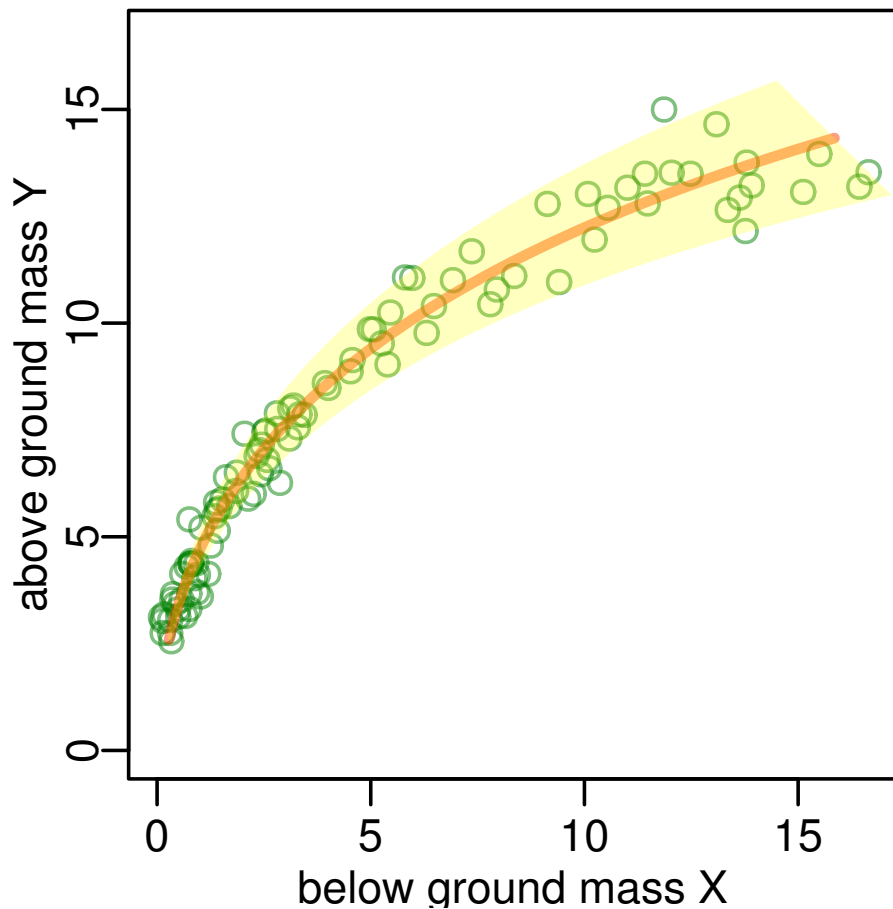
- 総重量 w 依存のパラメーター b はマイナス
- こういう問題は MCMC 収束が遅い

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**モデルではなかるーか?
- 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)

連続数量の分配モデルの応用例

- Iijima & Shibuya. 2010. J. For. Res 15:46–54.
- 樹木の枝内の重量分配 宮田さん (北大; 論文執筆中)
- 雌雄同株の植物個体内でのオス・メス繁殖器官への資源分配
- 植物個体内での資源分配を調べる安定同位体の存在比の解析
- 動物の行動観察記録に見られる「時間の分配」のモデル化..... ただし時系列構造の考慮も必要
- ほかにもいろいろあるかも?

今日のハナシのながれ

1. まずは $\{X, Y\}$ 誤差ありデータ解析における，ありがちな**回帰**適用お作法の問題点について検討
 - 因果関係なさそうなのに**無理に回帰**するのはヤメよう!
2. 解決策のひとつになりうる**重量分割モデル**の紹介
 - X と Y はどちらも結果だ
 - できればいくつかの X と Y の複数回測定を!
3. 重量分割モデルの拡張方法を検討
 - さらに生物学的な過程をとりこんだ改造も可能

— ベイズモデリングの自由さ —

アロメトリーとかありきたりな

方法が無批判に使うのではなく

モデリングの工夫も検討しましょう