「生態学基礎論 (生物多様性論 II)」の一部:

Statistical Modeling for Ecology, Jan 11, 2011

Part 2 in 2

An introduction of GLM

Better Data Analysis Using GLM

KUBO Takuya kubo@ees.hokudai.ac.jp

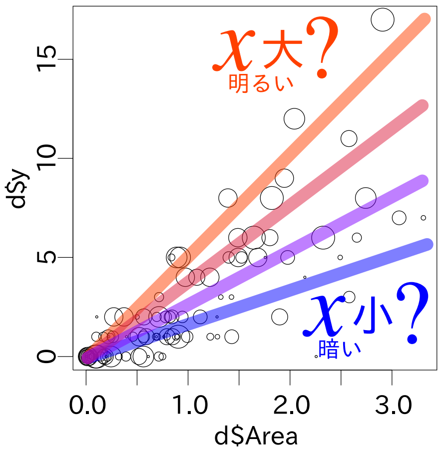
http://goo.gl/lqFgH

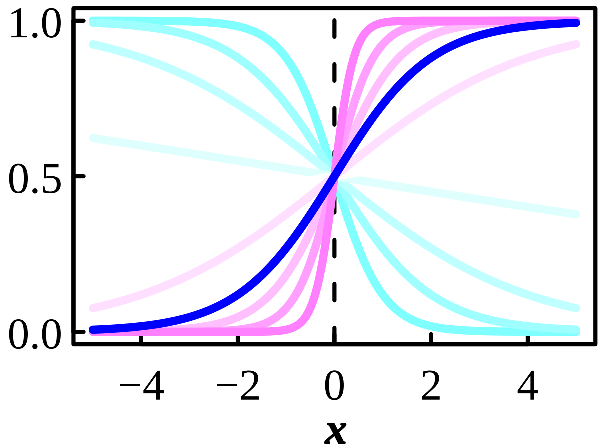
Statistical modeling using the generalises linear model (GLM)

1. **Modeling of observation** 1/17 (月)
   - What is statistical modeling? GLM?
   - Poisson regression, a part of GLM
2. **Stop the "Data / Data" manner!** 1/19 (水)
   - `offset` technique for Poisson regression
   - Logistic regression, as a part of GLM

Today's topics

1. Stop the "Data / Data" manner!
2. Enhancing Poisson regression with offset technique





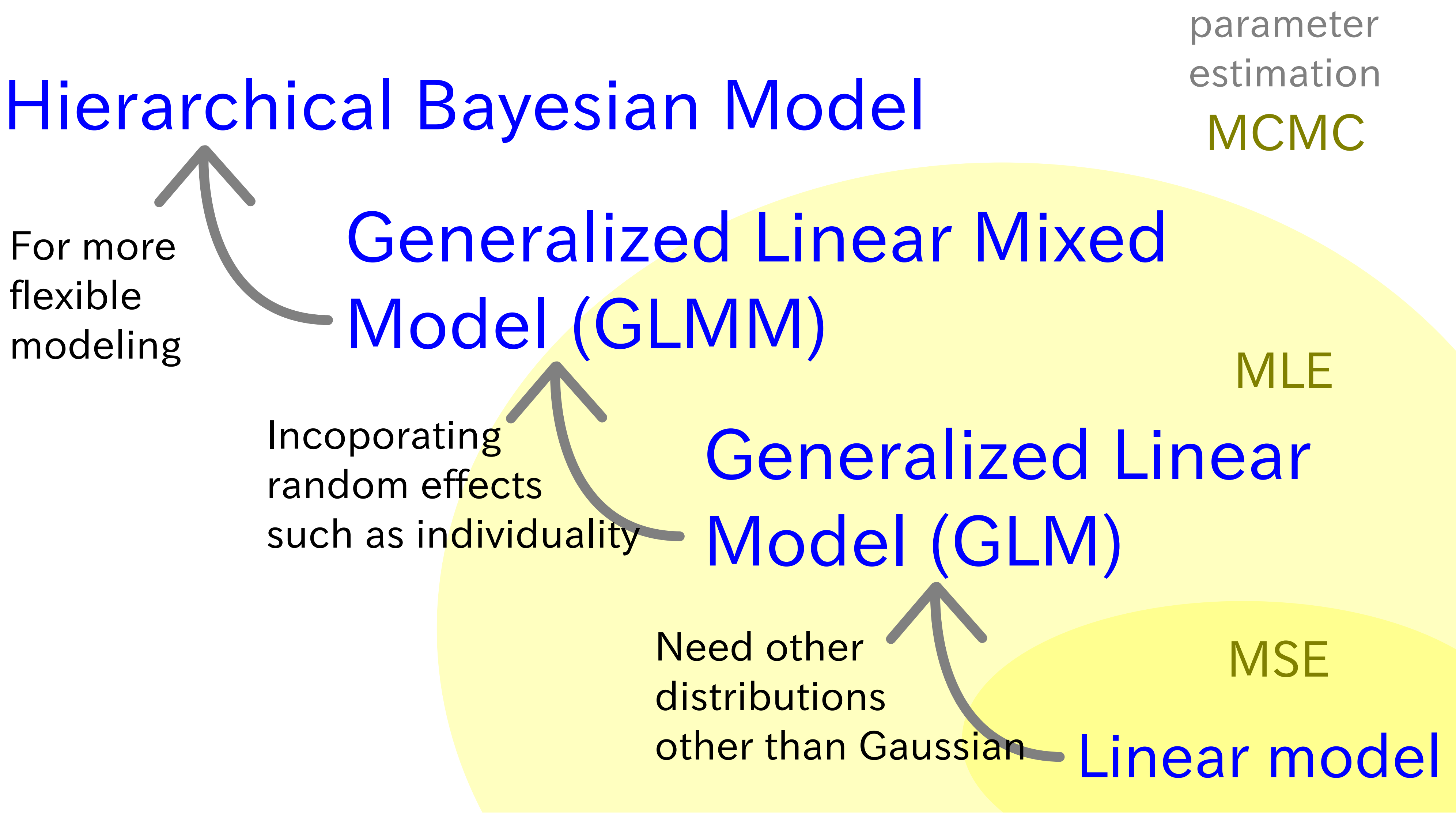
1. Stop the “Data / Data” manner

... plus, a short revision of the previous class

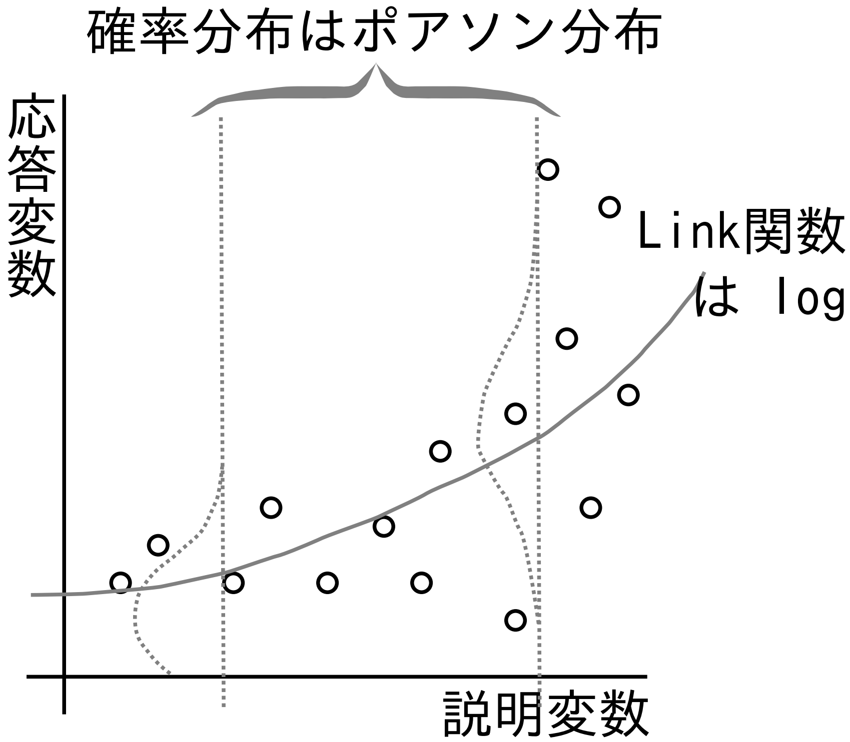
Statistical modeling of your observation

- Statistical modeling explains the patterns appeared in your data
- Probablistic distribution, the most important component of statistical model
- Goodness of fit to your data is evaluated by statistical models

The development of linear models



Poisson regression to represent patterns in "count data"



- Poisson regression is a part of GLM
- Variance of y depends on mean
- Non-negative model prediction

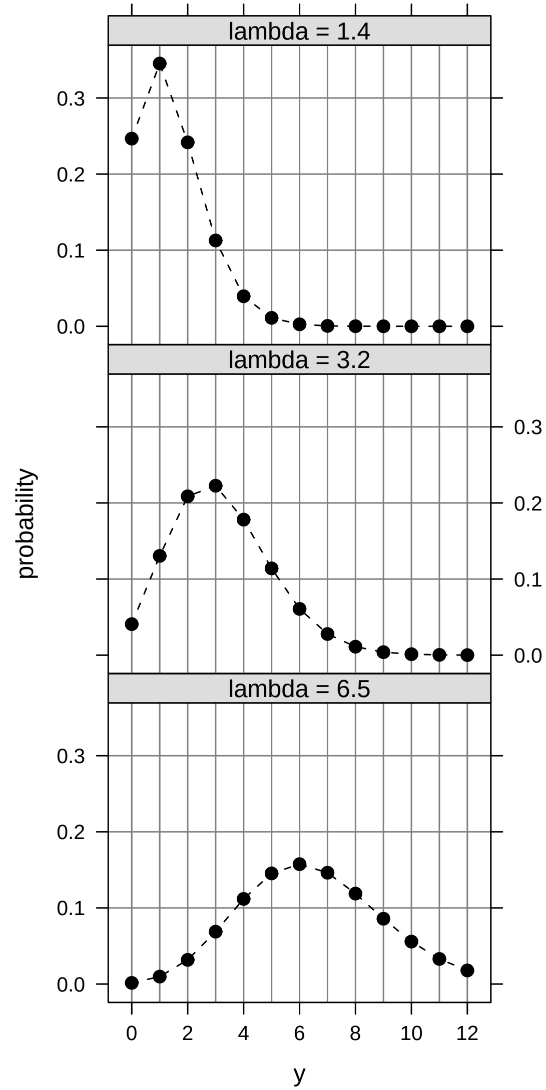
What is the Poisson distribution?

- A discrete probablistic distribution
- The functional form of Poisson distribution,

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

where λ is the mean of the distribution

- ... and the variance is equal to λ



- For discrete data (count data), unbounded
- e.g. egg number, seed number, population abundance ...

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

GLM can be specified by three components:

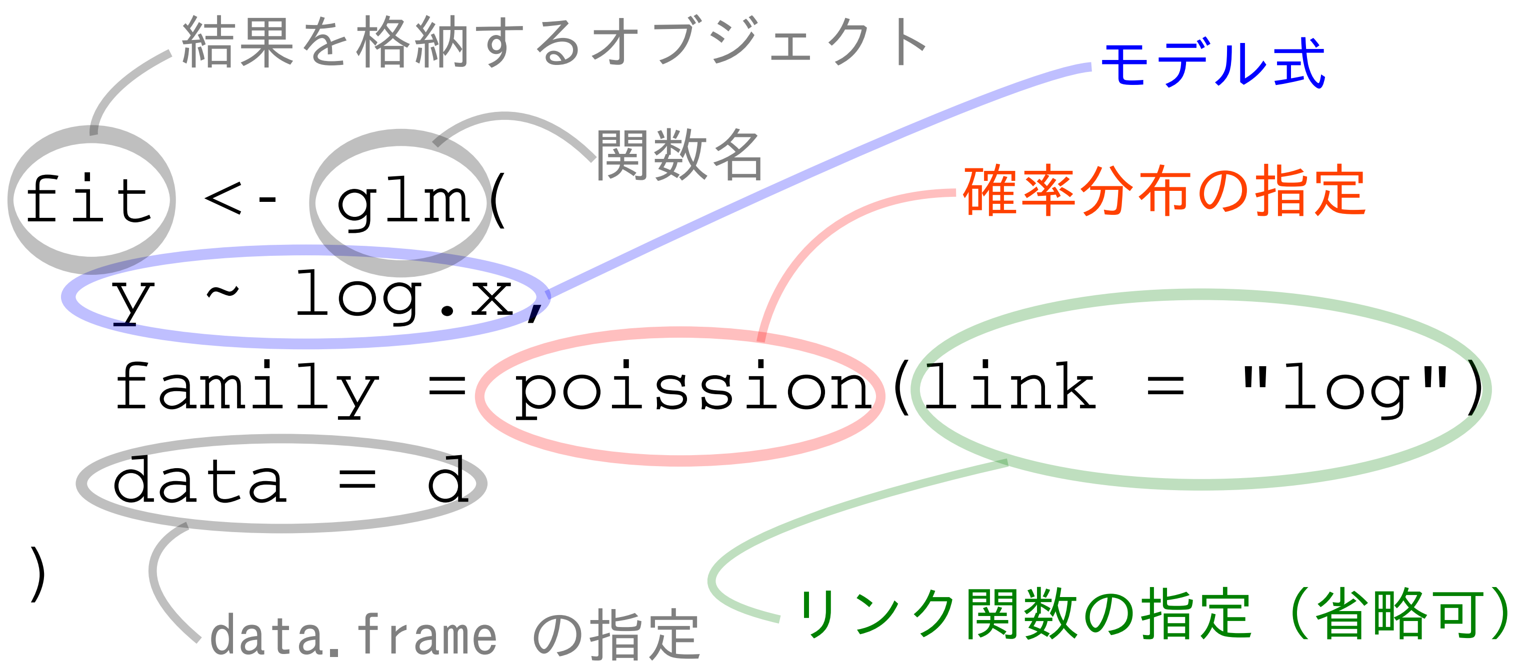
- Probablistic distribution: Gaussian, Poisson, Binomial, and other distributions
- Link function $f()$: (mean of y) $= f$(linear predictor)
- Linear predictor: $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$ where x_i and β_i are an explanatory variable and its coefficient, respectively
  - The coefficient set $\{\beta_i\}$ is estimated using the maximum likelihood method of which likelihood is defined by GLM and observed data

`glm()` function in R

	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)`

- some other `family` can be specified
- `glm.nb()` can be used by commanding `library(MASS)` in R

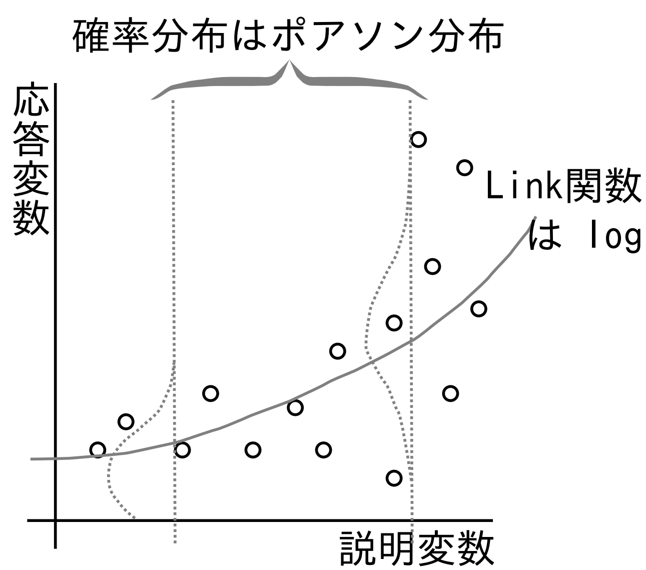
How do you specify the options of `glm()`



- model formula to specify response and explanatory variables
- `link` function: the functional form of y-mean
- `family` to represent the distribution of y

How to use `glm()` for Poisson regression

- `family`: `poisson`, Poisson distribution
- `link` function: `"log"` link function
- model formula: `y ~ x`



- **linear predictor** $z = a + bx$

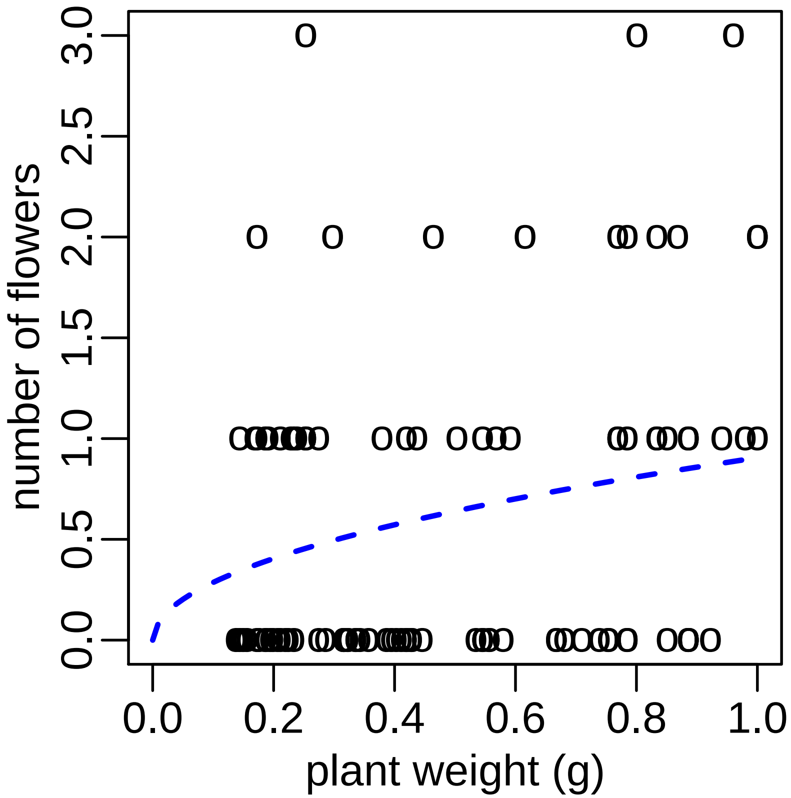
  both a and b are parameters to be estimated
- $\log(\lambda) = z$ where λ is the mean of y

  i.e., $\lambda = \exp(z) = \exp(a + bx)$
- response variable y follows the Poisson distribution of mean λ,

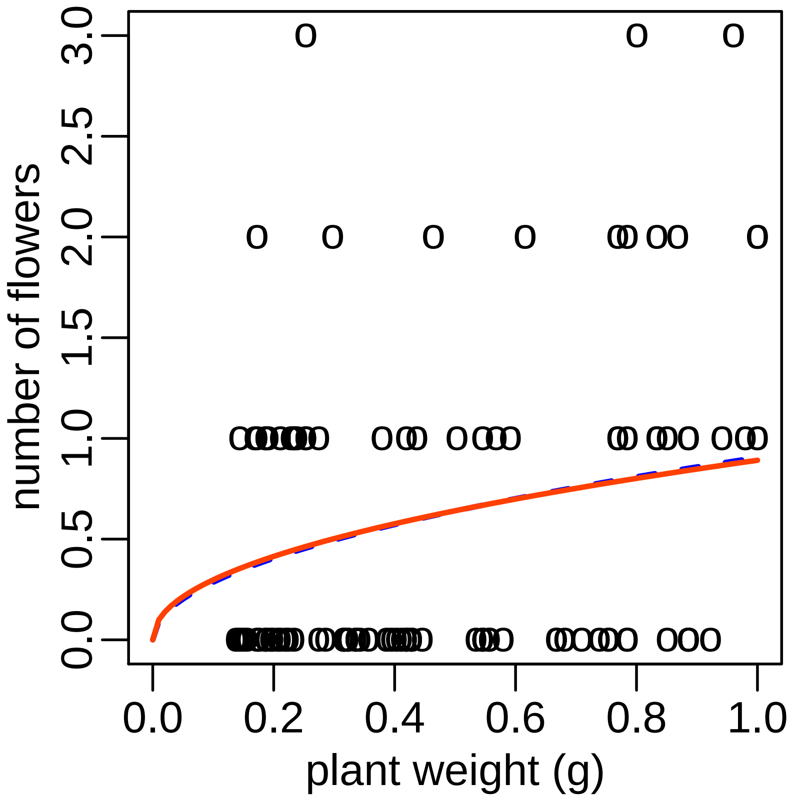
  $y \sim \text{Pois}(\lambda)$

Plotting the prediction based on GLM estimation

the “true” relationship between weight x and the number of flowers y



The estimated relationship



Some recommendations for data analysis

- Make many figures to show patterns in data
- Consider the probablistic distribution to represent data
- Stop “Data / Data” analysis! (next topic)

Cook your data without loosing its flavor and texture

How sad "Data / Data" analysis!

A frequently seen case in the unrecommendable manner ...

- You counted the number of flowering trees k_i in N_i trees in plot i
- You estimated the flowering probility p_i by evaluating k_i/N_i
- In plot j, $p_j = k_j/N_j$
- To know the "significant difference" betwenn p_i and p_j, you apply t-test in which you assumed p_* followed the Gaussian distribution ...

Why “Data / Data” analysis sucks?

- The distribution of **“Data / Data”**
- **Information collaption**: Is 3 / 10 and 60 / 200 same?
- Using statisitcal modeling, you no longer have reason to use “Data / Data” analysis
- No advantage in “Data / Data” analysis

How to stop "Data / Data" analysis?

- avoidable "Data / Data"
  - indices such as some densities

    e.g. population, wood densities

    escape tecnique: **offset term**
  - probabilities

    e.g. k items in N samples

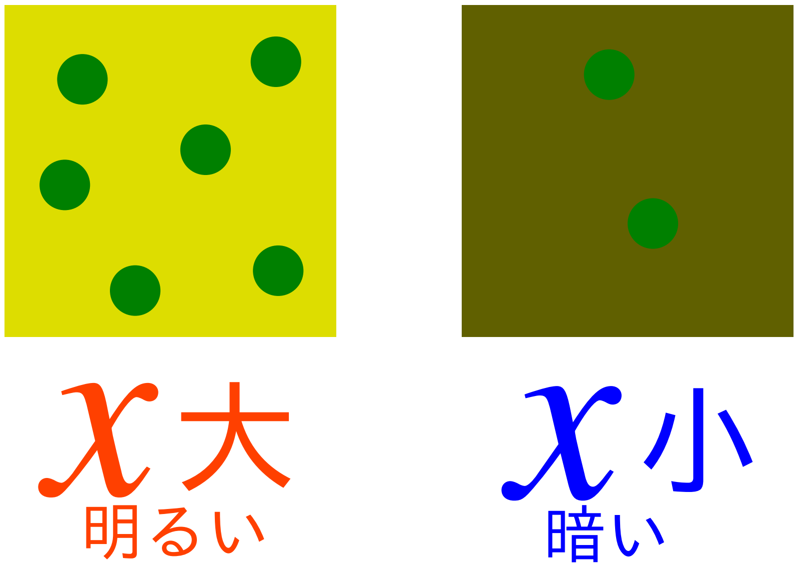
    escape tecnique: logistic regression, for example
- some quotients, hard to avoid
  - some measurement devices output fractions or densities ...
  - we sometimes make graphs using fractions ...

2. Offset Term Technique for Poisson Regression

to Stop "Data / Data" Analysis

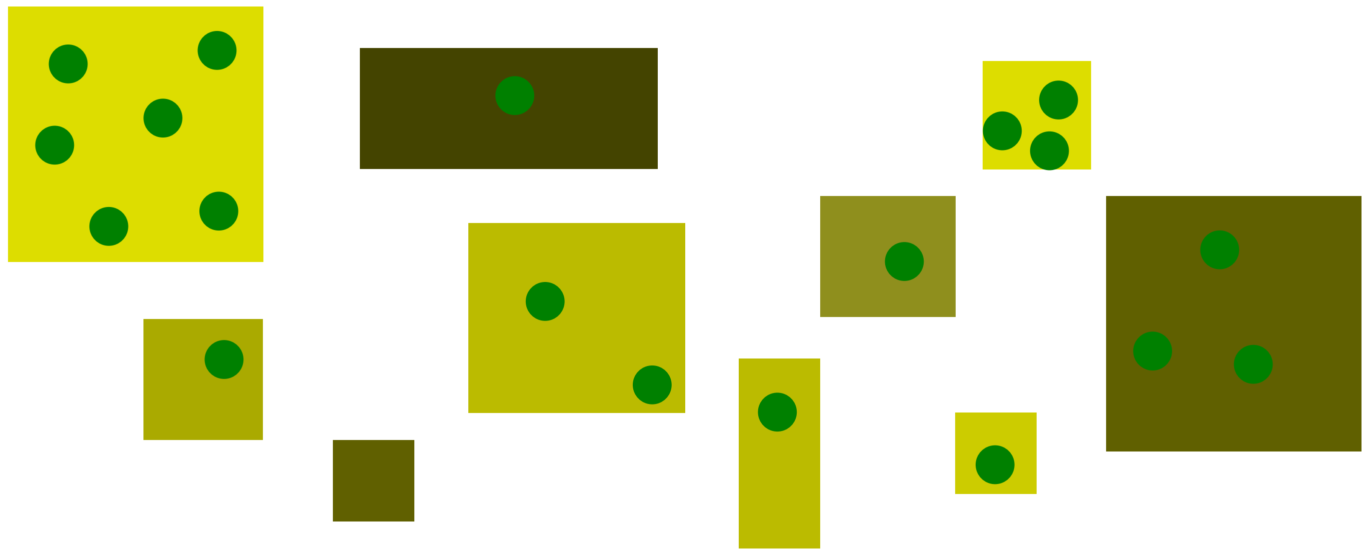
An example: density depending on light intensity

- To know the dependency of plant population density y on local light index x
- local light index $x \in \{0.1, 0.2, \cdots, 1.0\}$



Can we just apply `glm()` to estimate the effects of x?

What? Differences in plot size?!



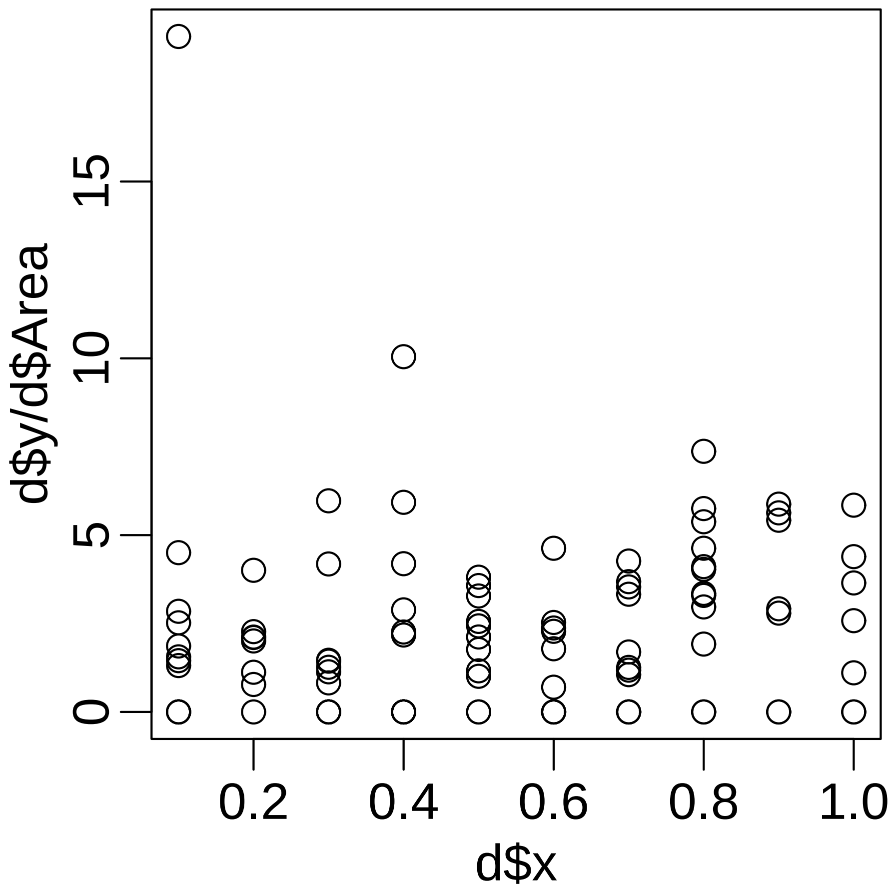
- We have to consider not only light index x but also plot area A
- Stop "density $= y/A$" estimation!
- We can manage it using `offset` technique for `glm()` function
- First, we have to draw figures of the data ...

data.frame in R: `Area`, light index `x`, plant abundance `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
```

light index x vs y/A

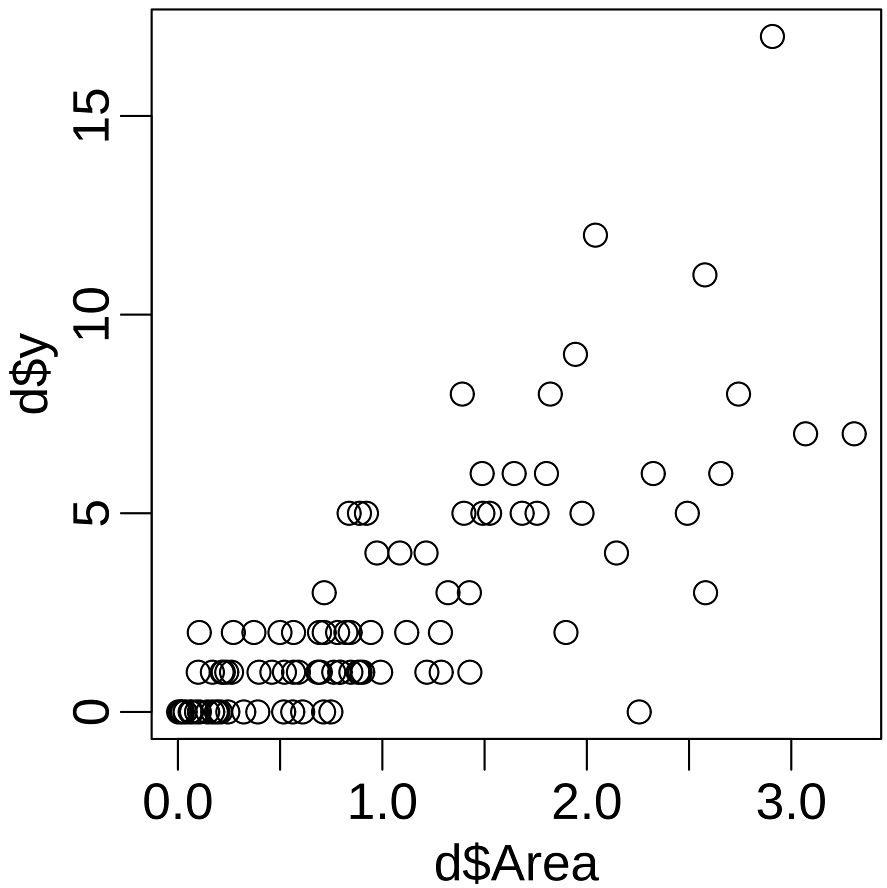
```
plot(d$x, d$y / d$Area)
```



- **UNCLEAR!**

Area A vs plant abundance y

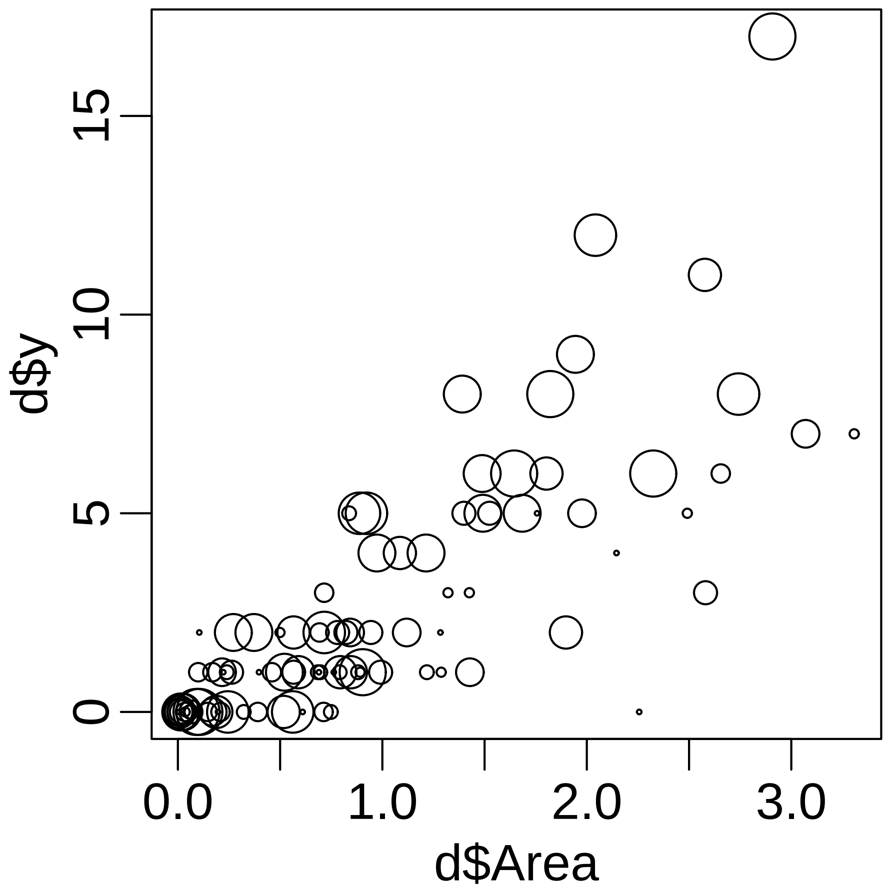
```
plot(d$Area, d$y)
```



- **Naturally, positvely correlated**

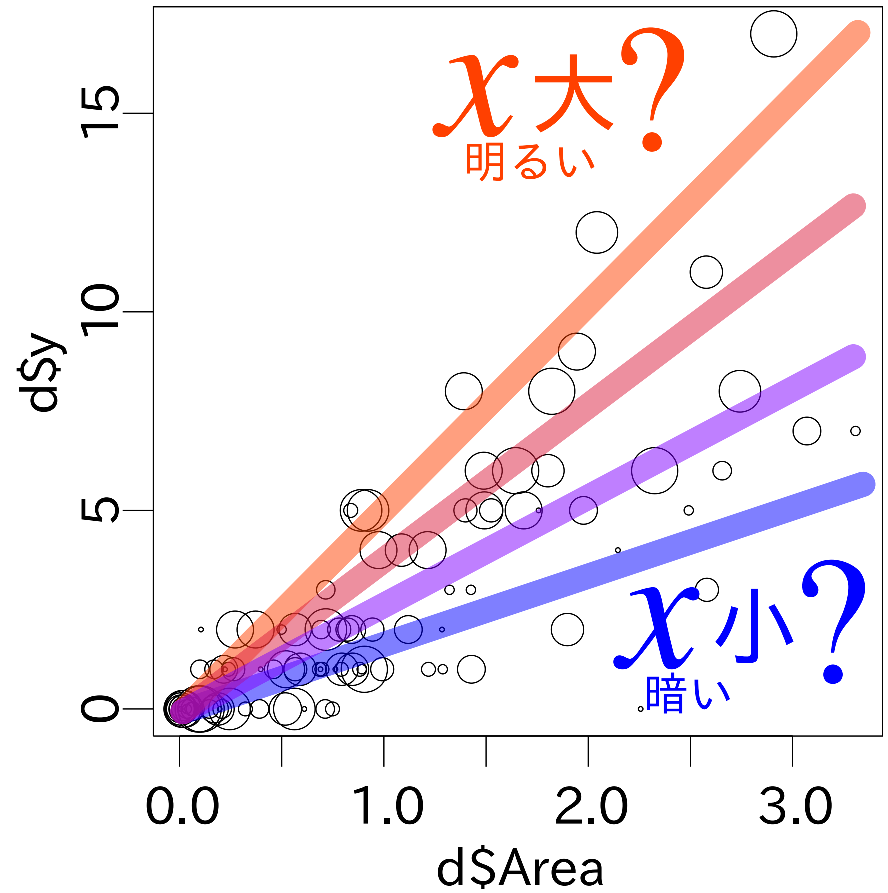
Adding x information by changing point size

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



- y increases with x when fixing A?

A statisitcal model in which plant density depends on x



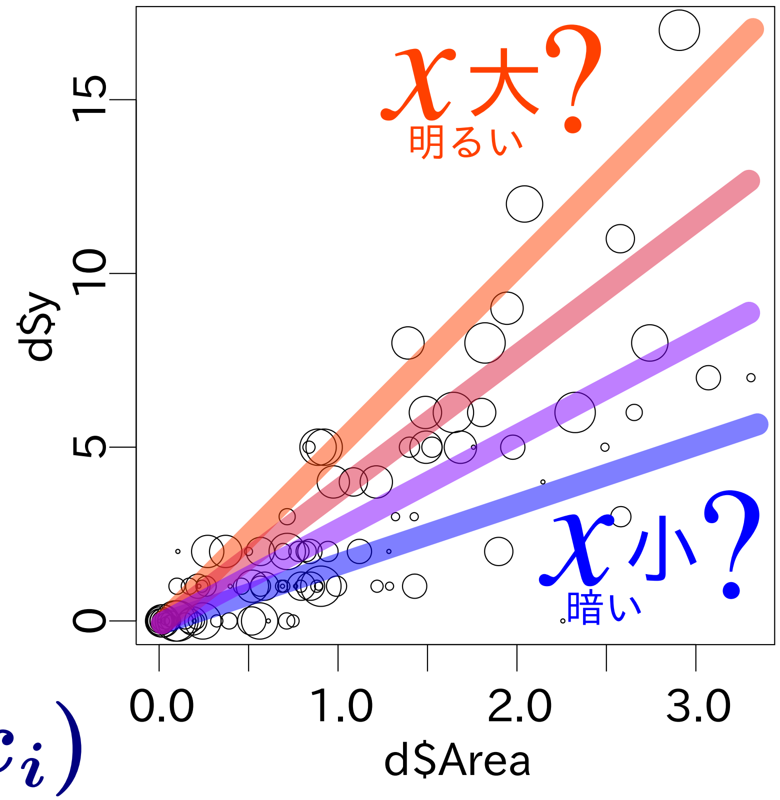
- the **mean** of population abundance is equal to $A \times$ (population density)
- population density depends on local light index x

Assumptions for the model

1. y_i follows the Poisson distribution of mean λ_i

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

2. $lambda_i$ is proportional to area A, and density depends on x_i



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

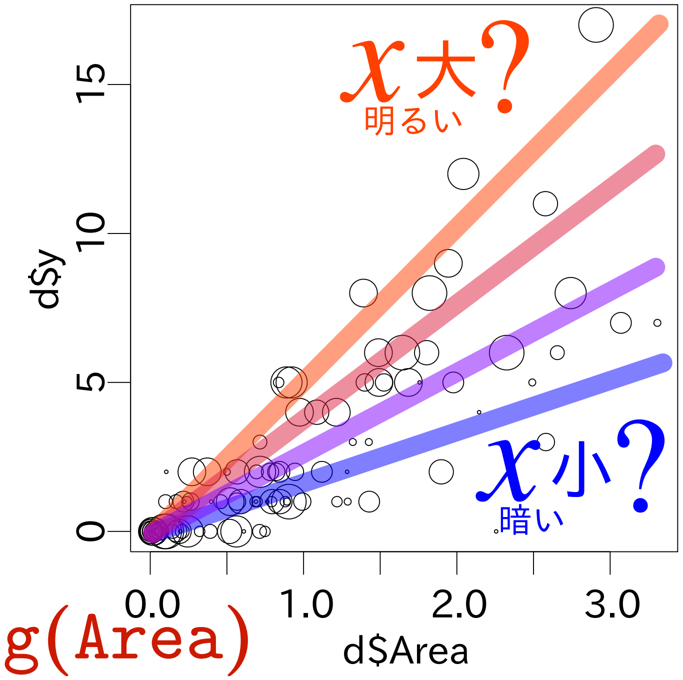
$$\lambda_i = \exp(a + bx_i + \log(A_i))$$

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

$\log(A_i)$ is referred to **offset term**

We can manage this mode using `glm()`!

- `family`: `poisson`, Poisson distribution
- `link` function: `"log"`
- model formula : `y ~ x`
- `offset` term: `log(Area)`



- **linear predictor** $z = a + b\,\mathrm{x} +$ `log(Area)`

  both a, b are parameters to be estimated
- $\log(\lambda) = z$ where λ is mean of y

  i.e., $\lambda = \exp(z) = \exp(a + b\,\mathrm{x} +$ `log(Area)`$)$

How to call `glm()`?

```
fit <- glm(
  y ~ x,
  family = poission(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

リンク関数の指定（省略可）

offset の指定

The estimated results using `glm()` of R

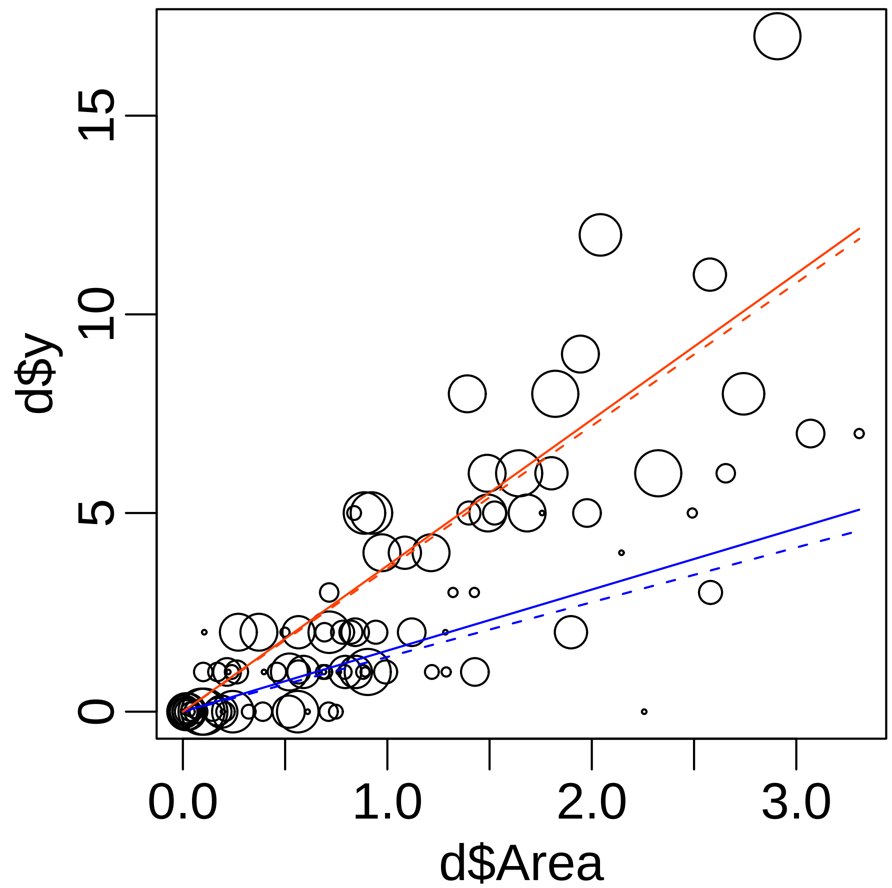
```
> fit <- glm(y ~ x, family = poisson(link = "log"), data = d,
  offset = log(Area))
> print(summary(fit))

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



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
```

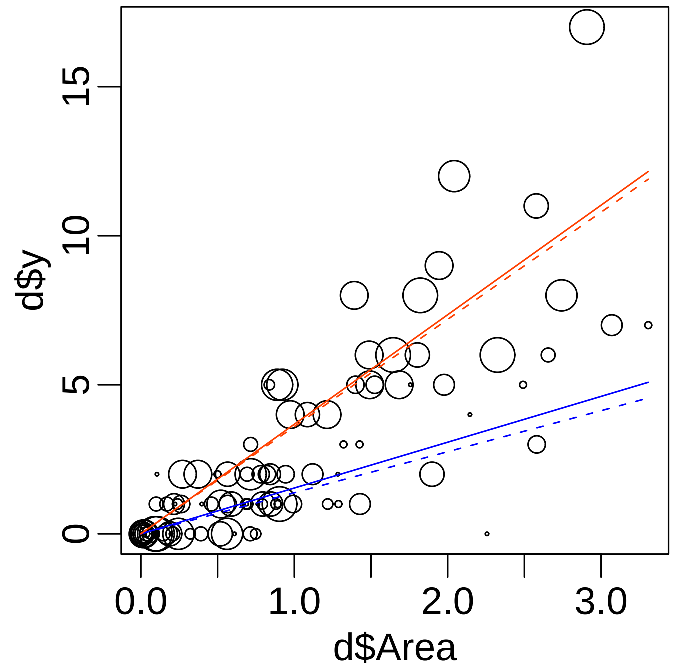
Plotting the model prediction based on the estimated results

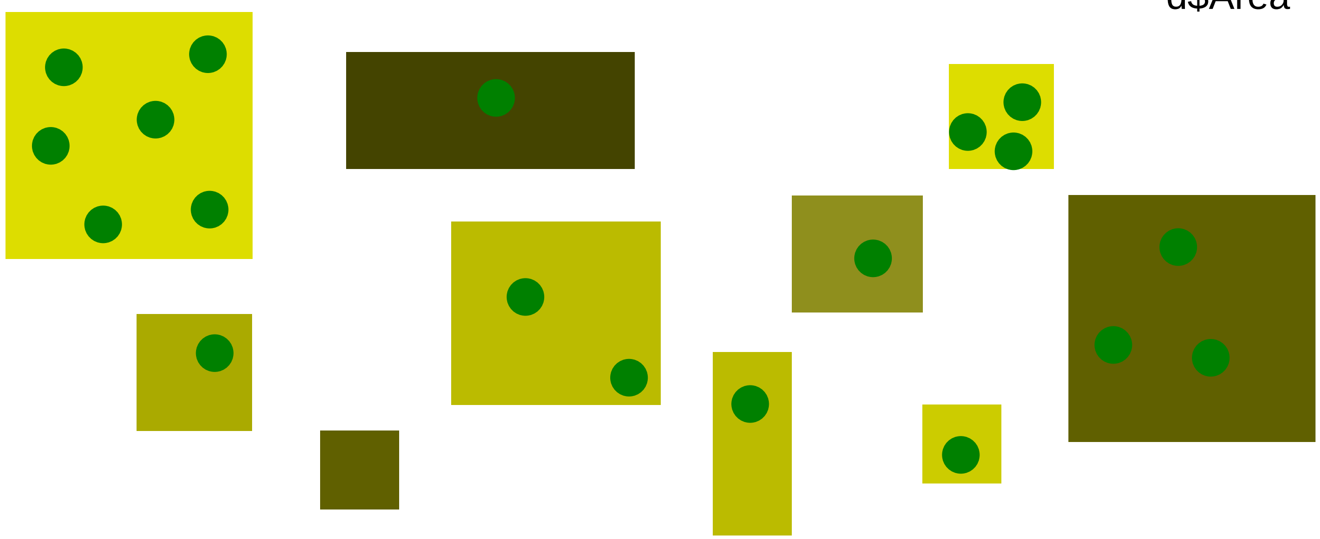


- solid red line for $x = 0.9$, blue for $x = 0.1$
- dashed lines are "true" line generating the example data

You can escape "Data / Data" analysis using `offset`

- In case that y is proportional to area A, $\log(A)$ must be specified as a offset term
- log(population density) is equal to linear predictor



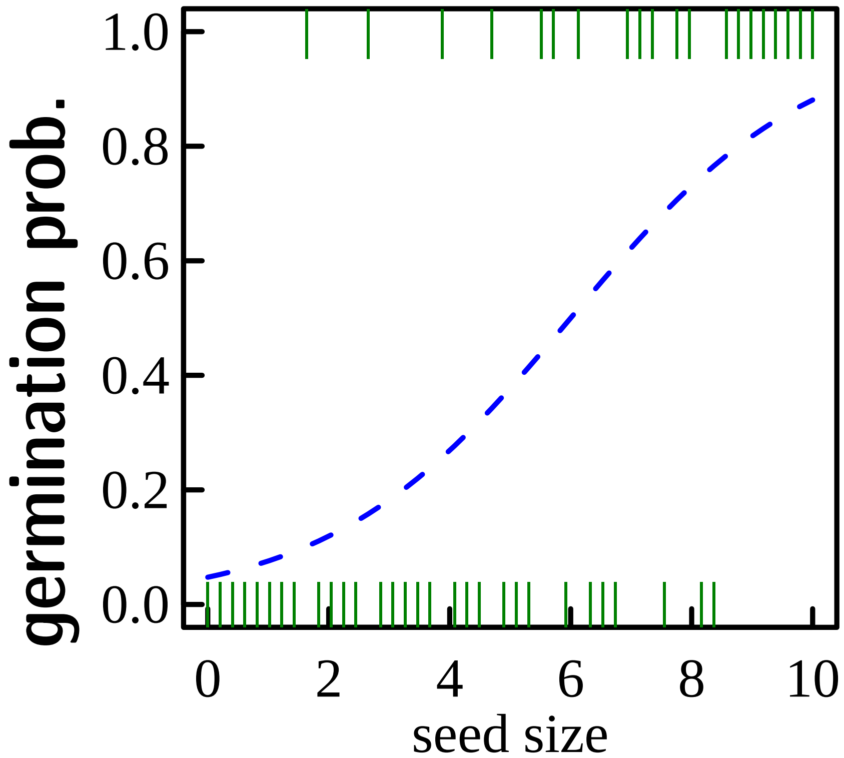


3. Logistic Regression

vs Unrecommendable Data Analysis

A fictitious example: germination data

Estimate the relationship between seed size and germination probability

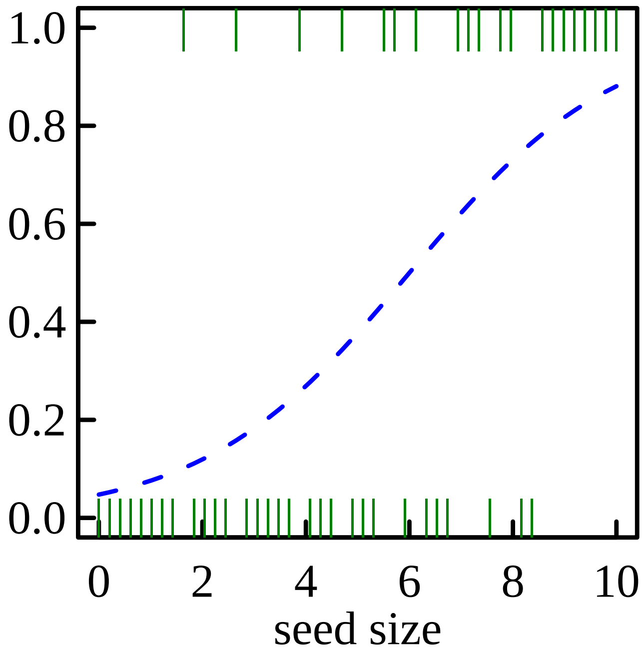


"true" germination curve

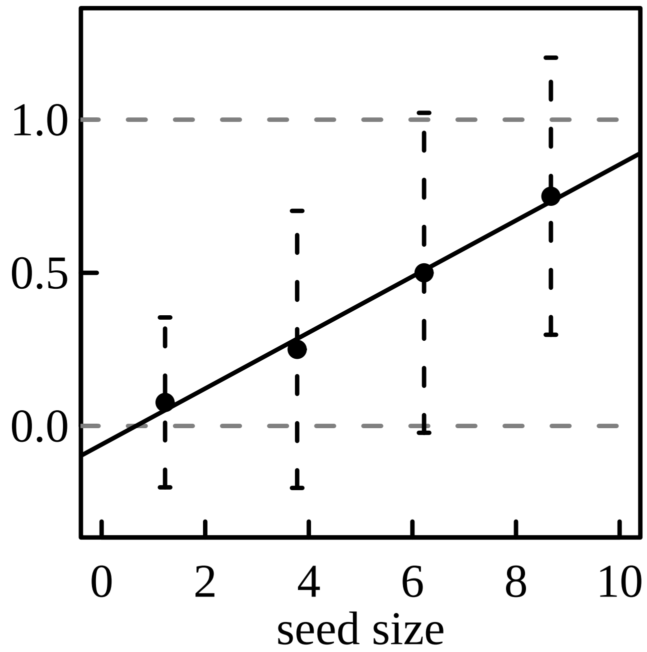
- germination prob. q increases with size x?
- How do we estimate q-curve (blue)?

Estimate "true" curve (in blue) based on the finite data

An unrecommendable analysis, but frequently seen ...



⇒

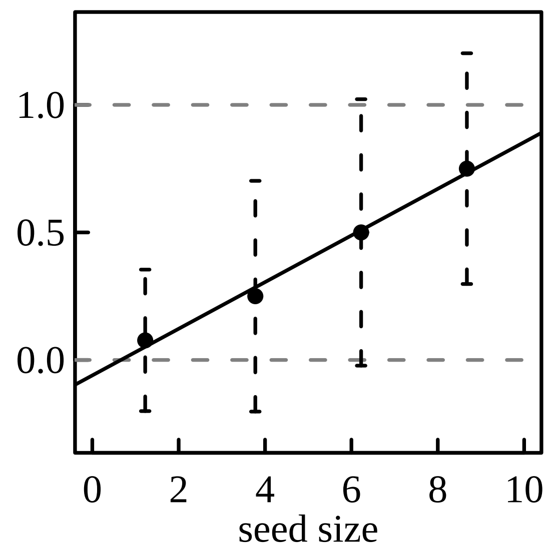


1. Data dividing and classifying along x-axis
2. Evaluating q for each size class using “Data / Data”
3. Throwing data into a black-box software

Why sucks? You neglected data characteristics

Arbitrary classification
Results depending on the arbitrariness

“Data / Data” erase information
Difference between 1 / 2 and 100 / 200!

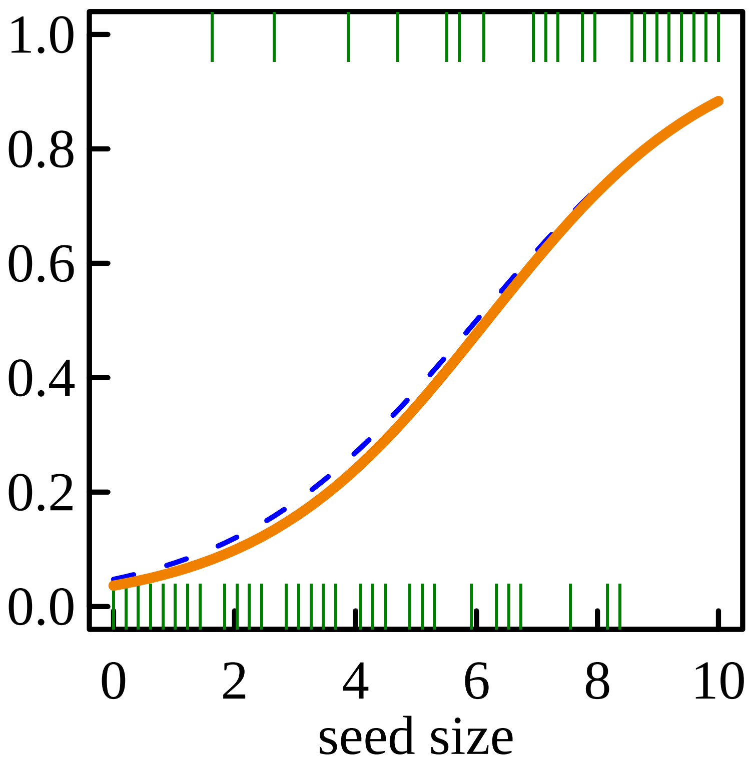


Neither normal nor homosedastic
therefore you can not apply any statistical models based on the normal (Gaussian) distribution

Surreal model prediction: germination prob. $q < 0$?!

Logistic regression using `glm()` function in `R`

Germination $y \in \{0, 1\}$ follows binomial distribution



- For each seed, germination prob. q is given as,

$$q = \frac{1}{1 + \exp(-(a + bx))}$$

(logistic equation)

- Using `glm()` function of R, we can estimate both a and b based on the given data

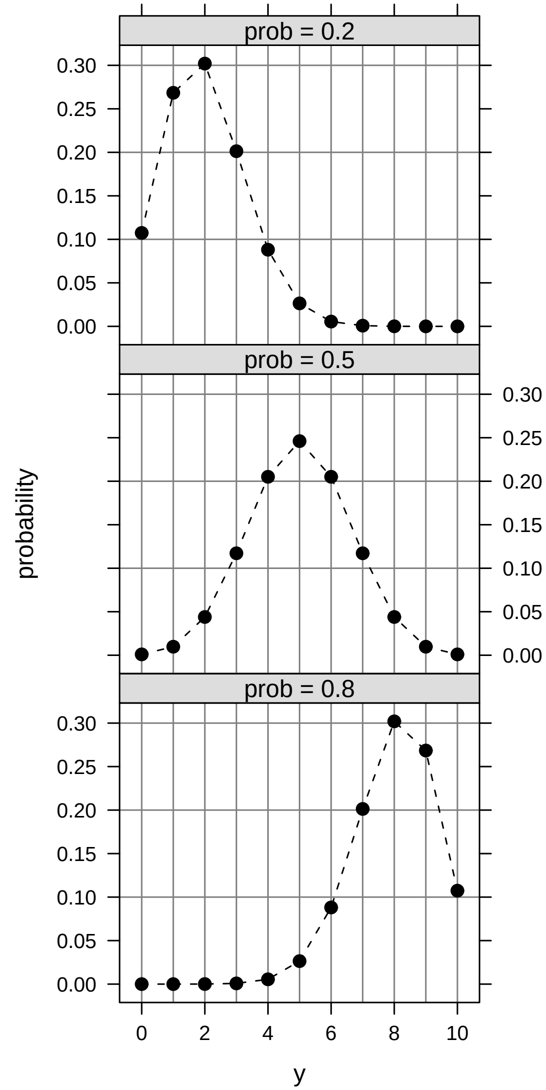
Binomial distribution

- Discrete random variable $y_i \in \{0, 1, 2, \cdots, N\}$
- 確率分布 (paramter: q, N) Probablistic distribution function:

$$\binom{N}{y} q^y (1-q)^{N-y}$$

- Mean Nq, Variance $Nq(1-q)$

- for upperbounded count data
- e.g., y individuals responded in size N population

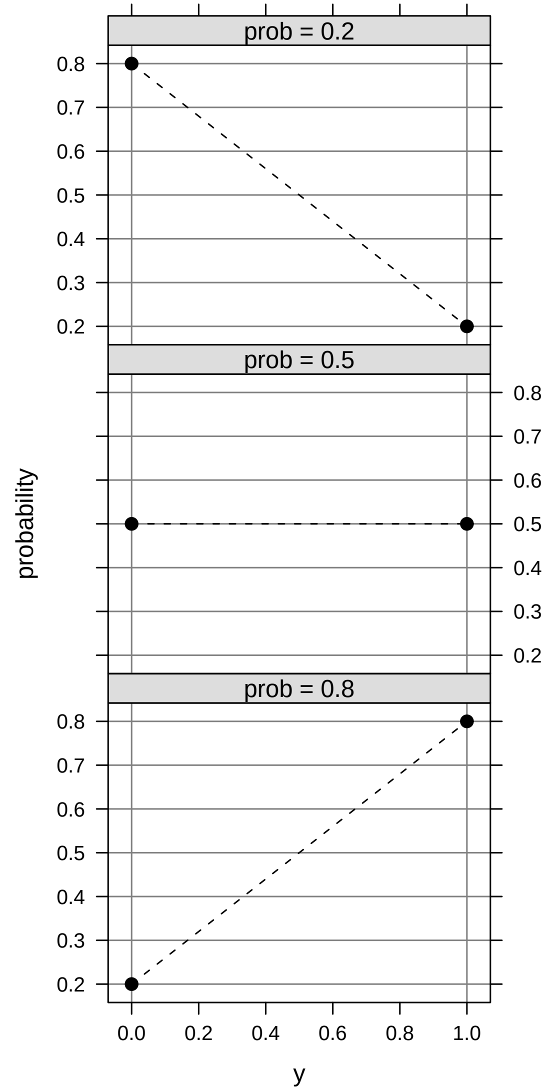


Bernoulli distribution

- Discrete random number $y_i \in \{0, 1\}$
- Probablistic distribution function:

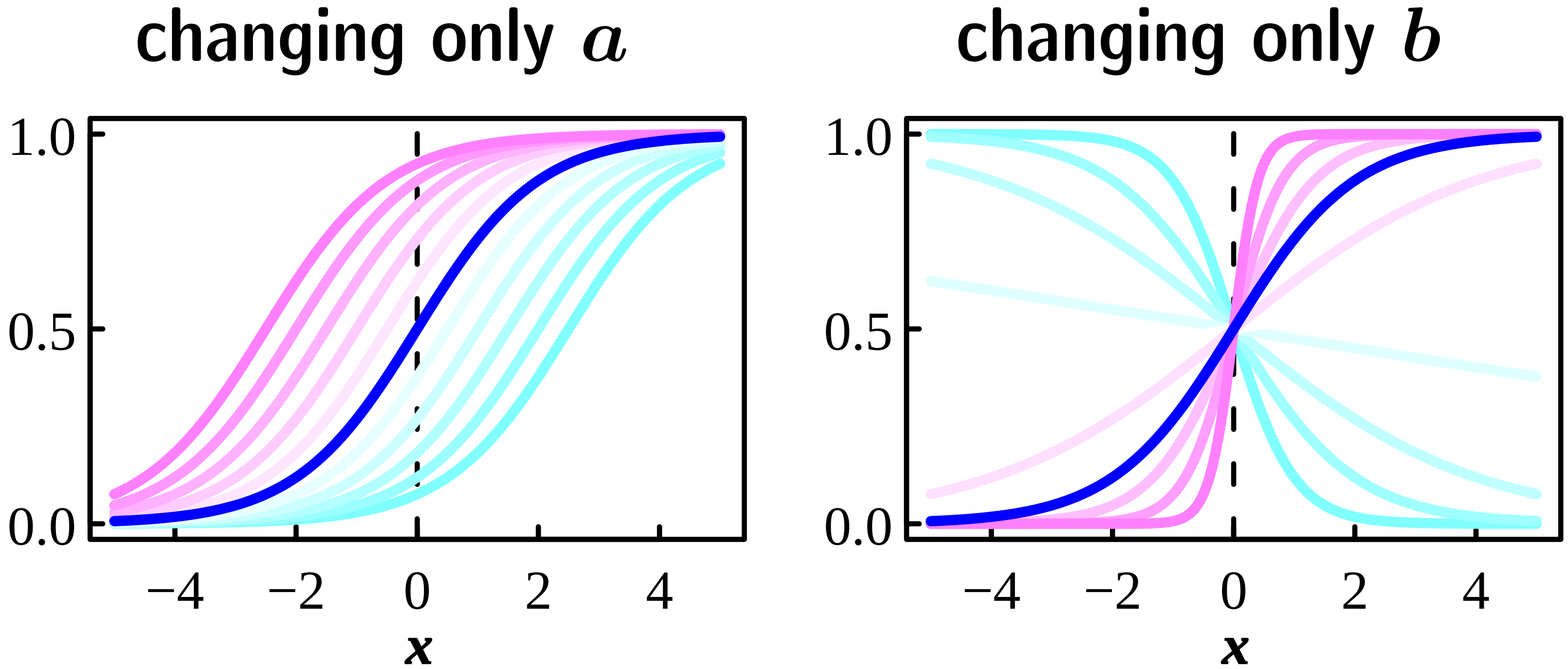
$$q^y(1-q)^{1-y}$$

- Mean q, Variance $q(1-q)$
- Bernoulli distribution is a special case when $N = 1$ in binomial distribution



Logistic function

$$q = \frac{1}{1 + \exp(-(a + bx))} \qquad (\exp(Z) = e^Z)$$



Variable q defined by a logistic function bounded in $0 \leq q \leq 1$

Logistic and logit functions

- logistic function

$$q = \frac{1}{1 + \exp(-(a + bx))} = \text{logistic}(a + bx)$$

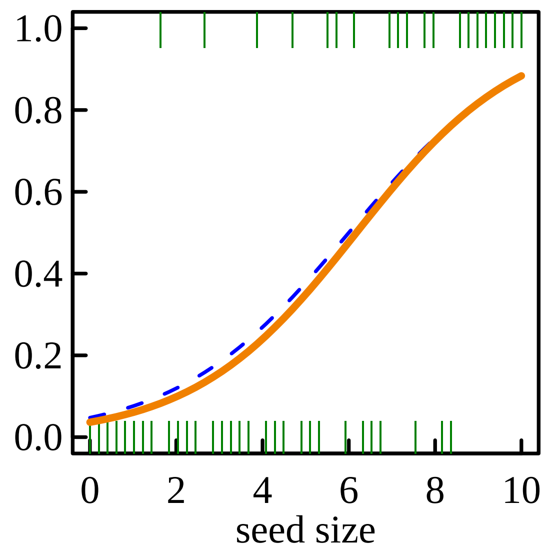
- logit transformation (logit function)

$$\text{logit}(q) = \log \frac{q}{1 - q} = a + bx$$

logit is the inverse function of logistic function, vice versa

Use `glm()` function for logistic regression (1)

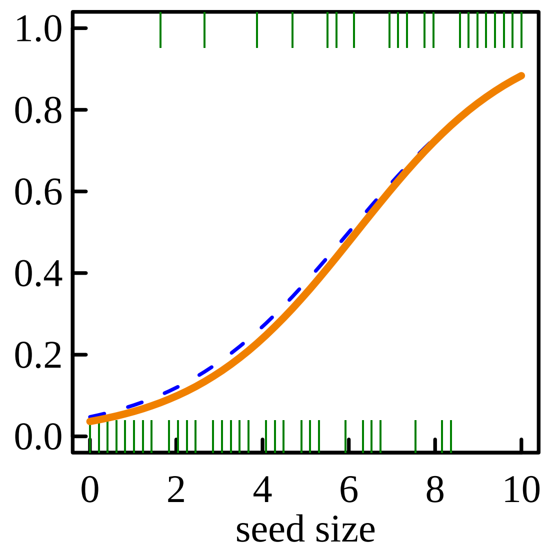
- `family`: `binomial`, 二項分布
  - $y \in \{0, 1, 2, \cdots, N\} \rightarrow$ binomial distribution
- `link` function: `"logit"`
  - `link = "logit"` is canonical under `family = binomial`
- model formula: `y ~ x`



What is represented by `family = binomial(link = "logit")`

Use `glm()` function for logistic regression (2)

- `family`: `binomial`, binomial distribution
- `link` function: `"logit"`
- model formula: `y ~ x`



- linear predictor $z = a + bx$

  both a and b are pramters to be estimated based on data

- the relationship between germination probability q and seed size x,

$$q = \frac{1}{\exp(-z)} = \frac{1}{1 + \exp(-(a + bx))}$$

- response variable y follows ...

  $y \sim \text{Binom}(q, N)$

In R, glm() must be specified as,

```
fit <- glm(
  y ~ x,
  family = binomial(link = "logit"),
  data = d
)
```

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

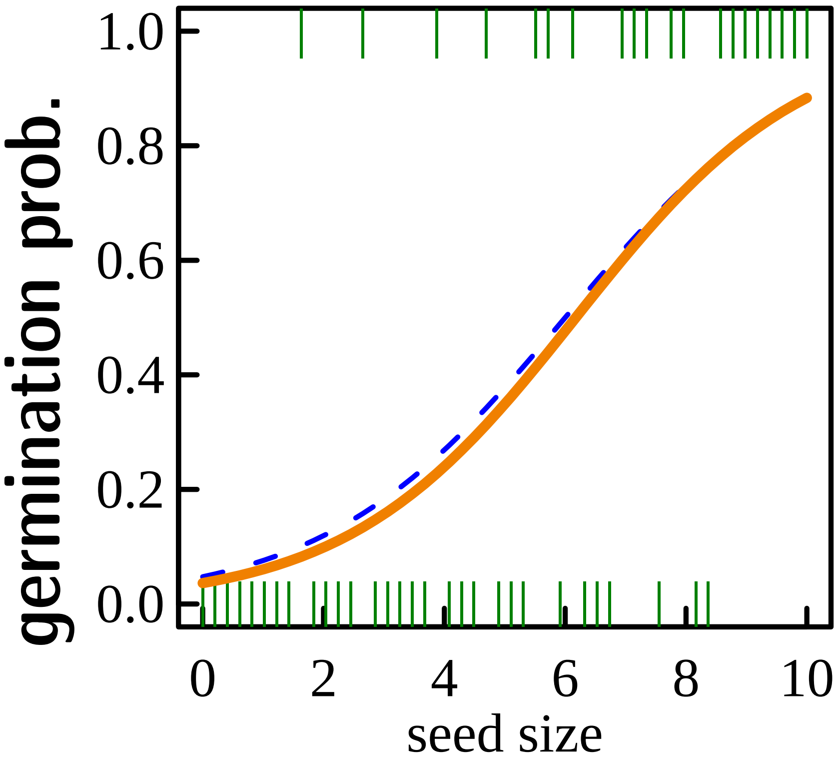
リンク関数の指定（省略可）

data.frame の指定

- model formula: seed size x, explanatory variable
- link function: logit
- family: binomial, binomial distribution

Ending

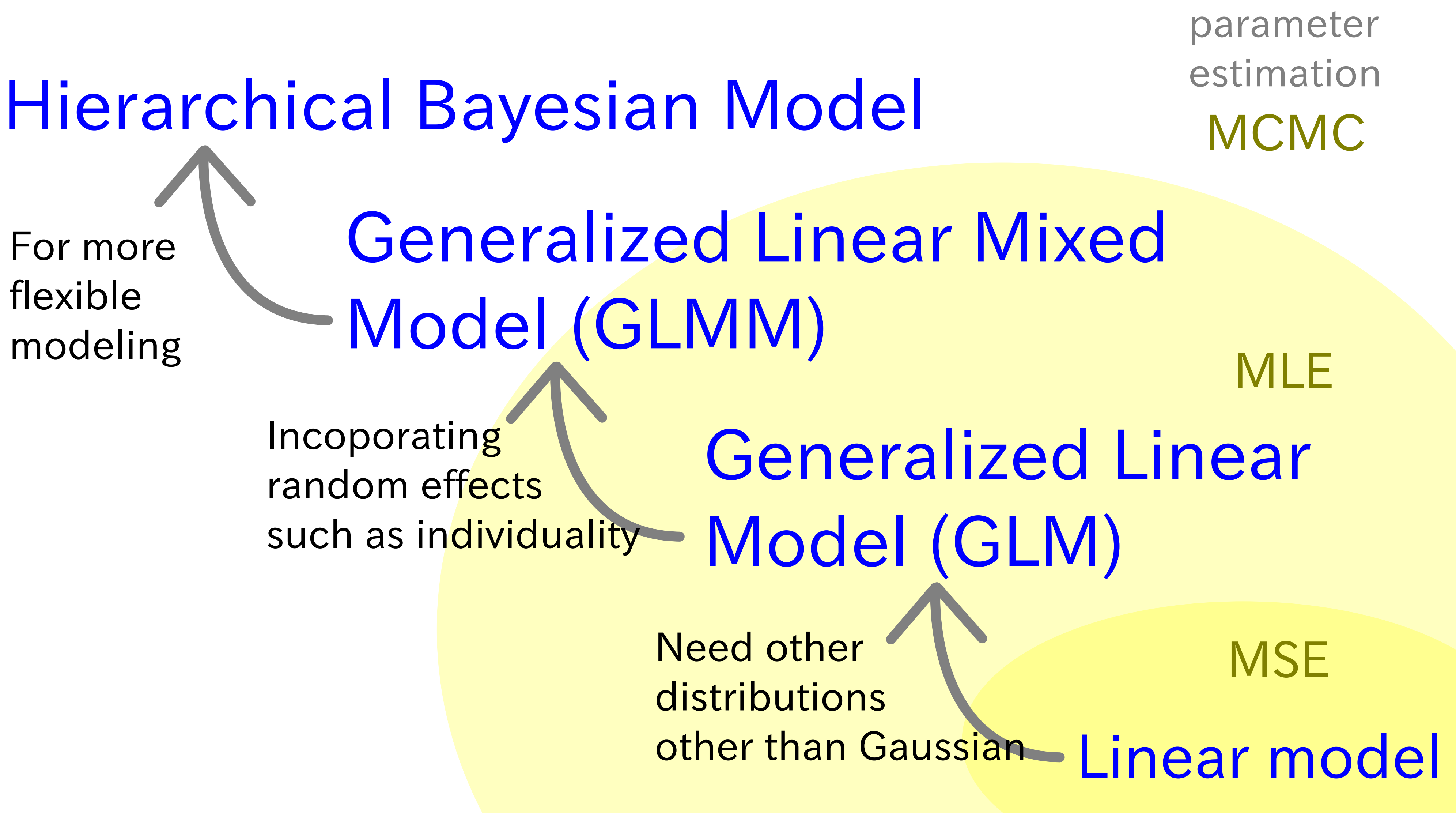
Closing: for your better data analysis



- Don't divide data arbitrarily
- No "Data / Data" analysis
- Plot your data in several way as many as you can
- Seek the best probablistic distribution to represent your data

Conclusion: Don't overcook your data, look at the natural aspect of your data

The development of linear models



A learning plan: development of GLM family