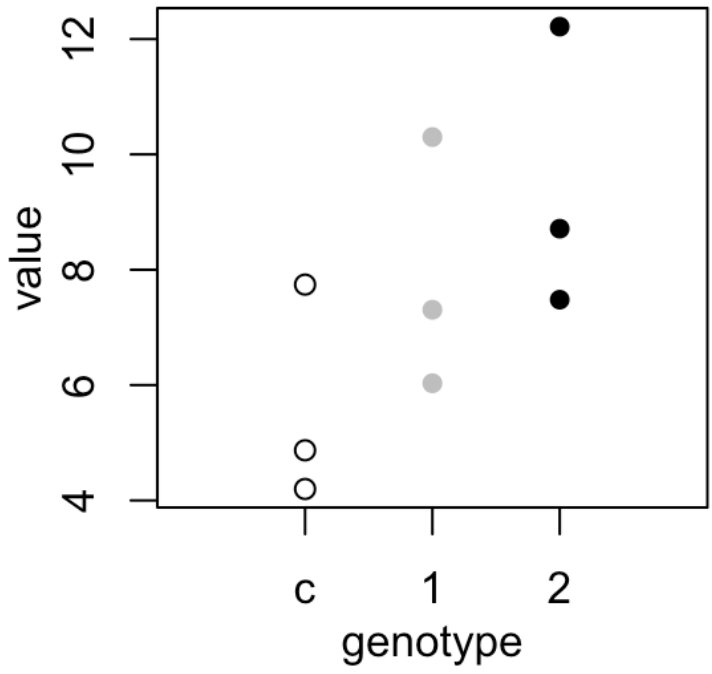
線形モデル・計画行列

北海道大学大学院農学院
（兼）数理・データサイエンス教育研究センター
佐藤昌直

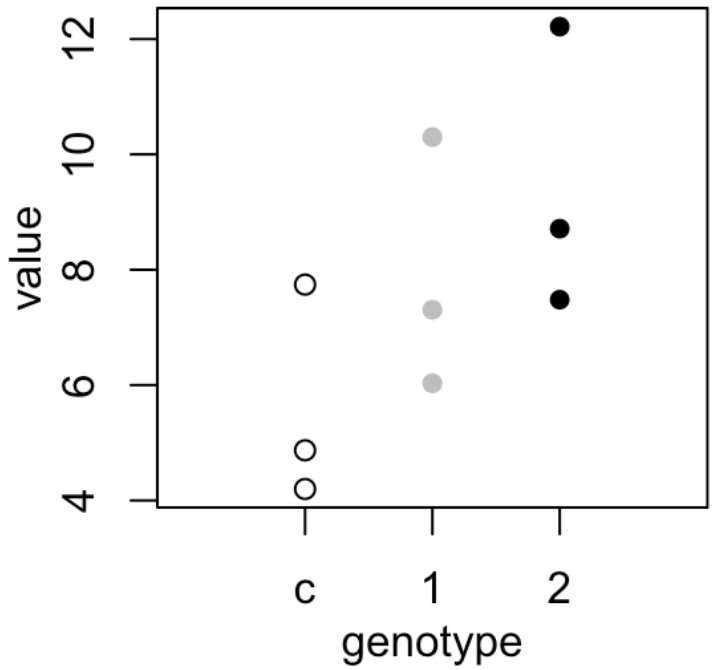
あるRT-qPCR実験: 生データ



genotype
- control
- strain1
- strain2

replicate: 1, 2, 3

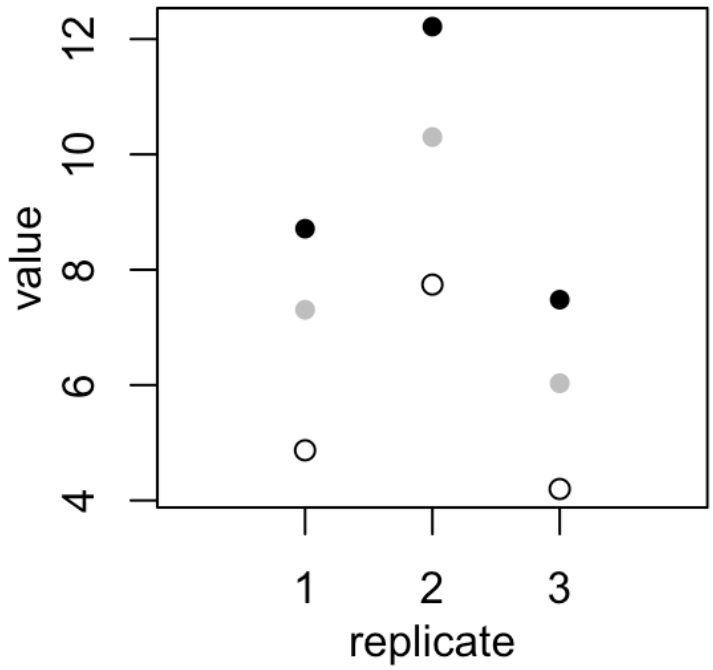
あるRT-qPCR実験: t検定結果



p-values

- control vs strain1 = 0.2456
- control vs strain2 = 0.1011

生データをreplicateについて可視化



genotype

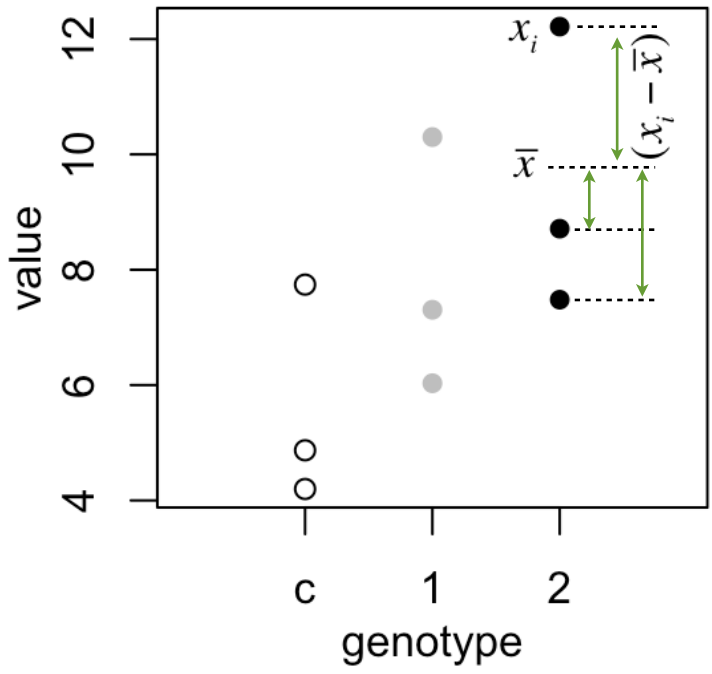
- control
- strain1
- strain2

replicate: 1, 2, 3

ポイント

検定から推定（予測・モデル構築）へ：線形モデルへの転換

線形モデルで考えてみる：モデル表記



$$x_1 = \bar{x} + (x_1 - \bar{x})$$

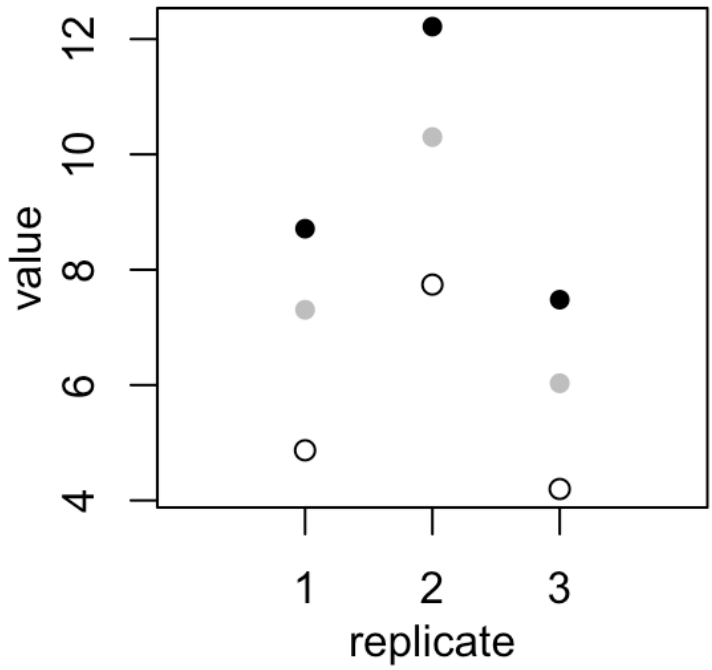
線形モデルで考えてみる：モデル表記

$$x_1 = \bar{x} + (x_1 - \bar{x})$$

$$x_1 = \bar{x} + \varepsilon_1$$

残差 (観察値-推定値):
想定要因では説明できない
データの変動

生データをreplicateについて可視化



genotype

- control
- strain1
- strain2

replicate: 1, 2, 3

観察値を複数要因の
影響に起因するものとして分解

$$x_1 = \bar{x} + (x_1 - \bar{x})$$

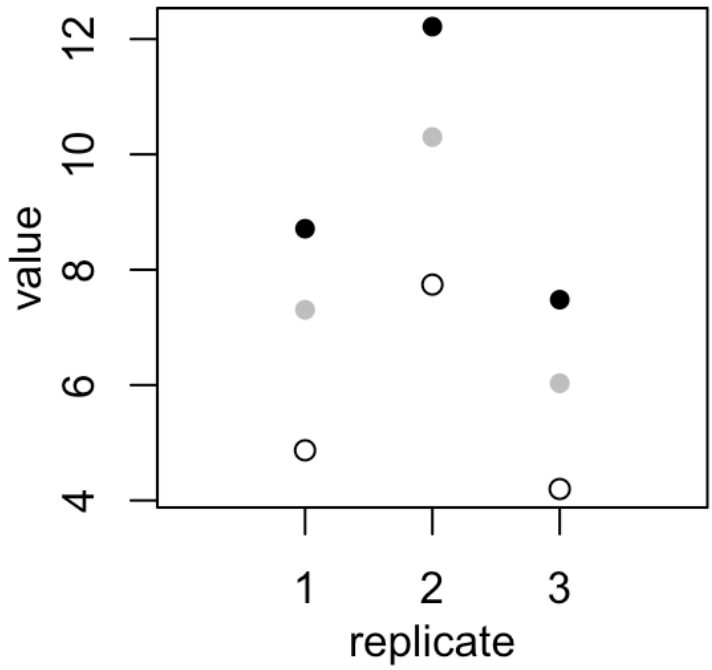
$$x_1 = \bar{x} + \varepsilon_1$$

↓

$$O_{ij} = x_i + y_j + \varepsilon_{ij}$$

*genotype*と*replicate*の
影響を同時に
考えられないか？

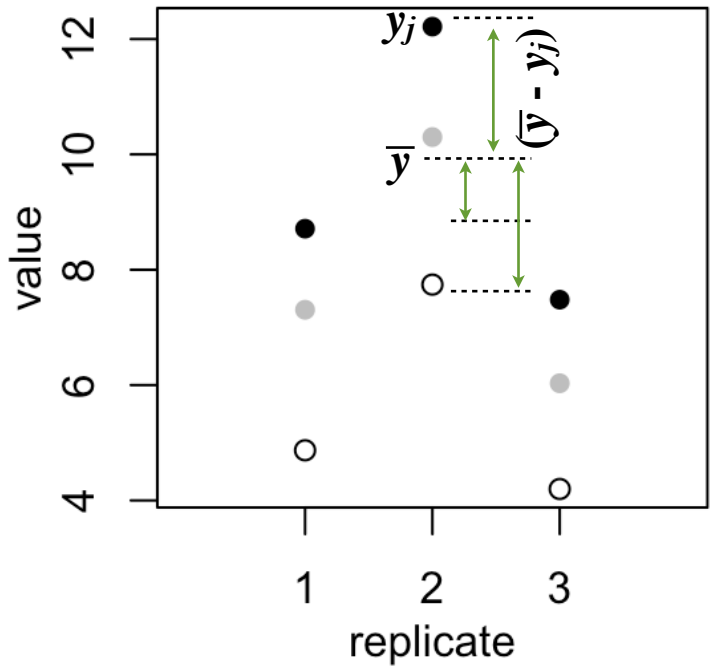
生データをreplicateについて可視化



genotype
- control
- strain1
- strain2

replicate: 1, 2, 3

replicateの影響も推定

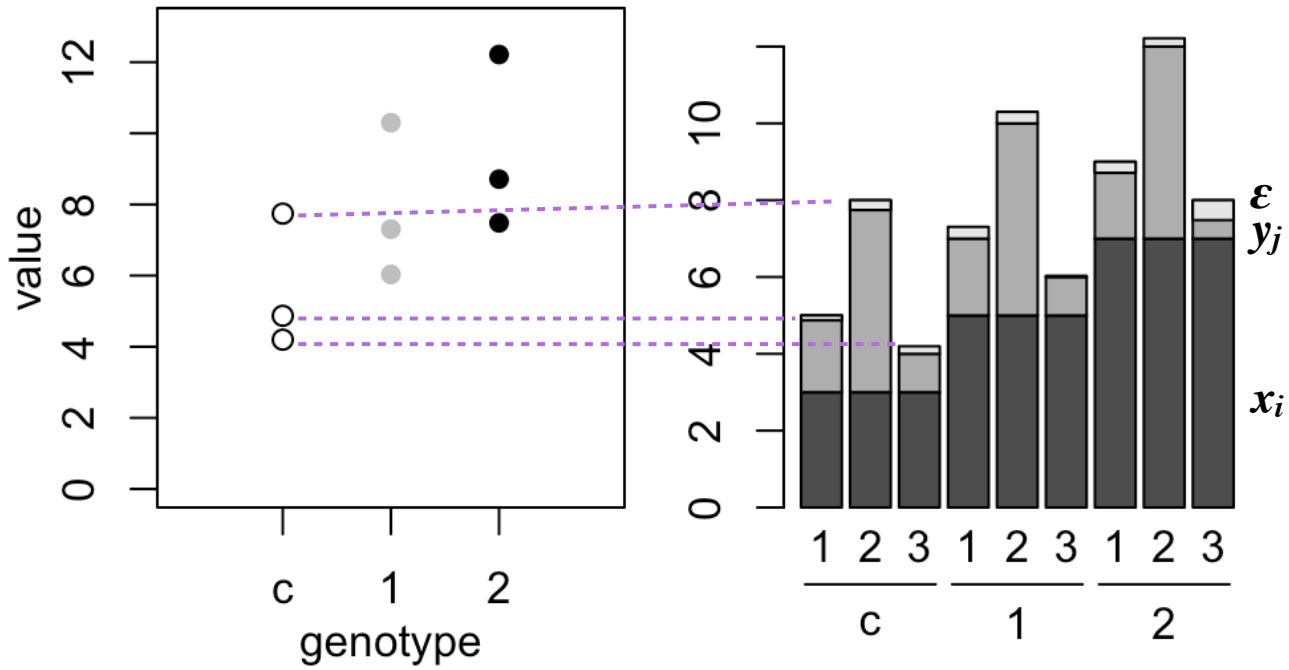


genotype

- control
- strain1
- strain2

replicate: 1, 2, 3

genotype, replicateの影響を同時に推定する: $O_{ij} = x_i + y_j + \varepsilon_{ij}$



線形モデルの枠組み

$$O_{ij} = x_i + y_j + \varepsilon_{ij}$$

教科書・論文での書き方

$$O_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

応答変数　　説明変数

μ: 切片(intercept).

回帰分析 $y=ax+by+c$のc.

Rのモデル式では明示的には

含める必要がない.

線形モデルとは

応答変数　～　説明変数1 + 説明変数2 + ... + 誤差

観察値（応答変数）を説明する（かもしれない）**説明変数の足し算**で応答変数の増減への貢献度を推定する

- R: `lm`, `glm`, `glmFit`などの関数を使う

実験デザインの重要性

- -omicsデータは“**batch effect**”と呼ばれる体系的なバイアスが混入する。

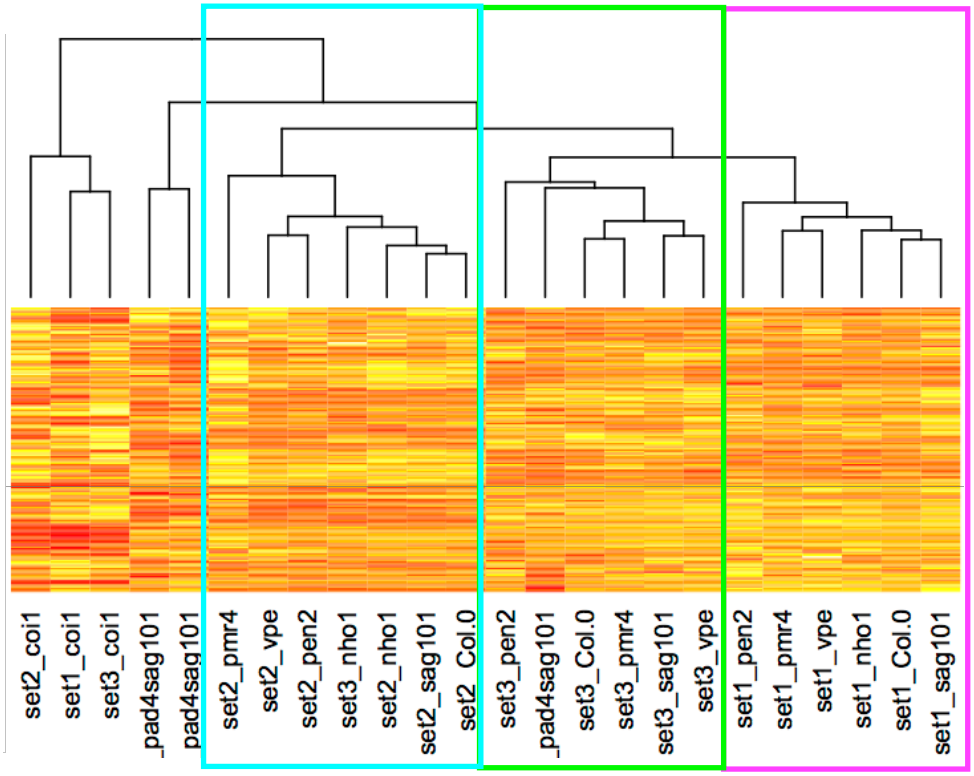
  例: 実験時期、実験者、餌

Nature Reviews Genetics (2010) 11, 733-

- 線形モデルで推定・除去

ポイント

batch effect の トランスクリプトームへの影響



ポイント

実験デザインの重要性

- 線形モデルで推定・除去

$$O_{ij} = \alpha_i + \beta_j + \varepsilon_{ij}$$

α_i : 遺伝子型／処理など注目している効果の要因

β_j : 反復（実験日時）／実験者などバイアス要因

- α_i の推定値、標準誤差のみを使う

R (egdeR) での実装: カテゴリカル因子とモデル式

```
(example)

$ R
> library(edgeR)                          #load edgeR library
> dat <- read.delim("count_data.txt", …) #import count table to R

> treat <- factor(c("M", "M", "M", "H", "H", "H"))   カテゴリカル因子
> treat <- relevel(treat, ref="M")
> design <- model.matrix(~treat)                     モデル式
> rownames(design) <- colnames(y)

> D <- DGEList(dat, group=treat)      #import data to edgeR
> D <- calcNormFactors(D, method="TMM")   #normalization (TMM)
> D <- estimateDisp(D, design)        #estimate dispersion
> fit <- glmFit(D, design)            #fitting to model
> lrt <- glmLRTt(D, coef=2))          #DE test
> topTags(lrt)
> …
```

ポイント

R (egdeR) での実装:
カテゴリカル因子とモデル式

$$O_{ij} = x_i + y_j + \varepsilon_{ij}$$

↓ 教科書・論文での書き方

$$O_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

応答変数　　説明変数

↓ Rでの書き方

```
o ~ a + b
```

(model.matrix関数の場合： ~ a + b)

model.matrixでの切片

$$O_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

応答変数　　説明変数

μ: 切片(intercept).

回帰分析 $y=ax+by+c$のc.

Rのモデル式では明示的に

示す場合に1を、切片を

含めない場合には-1を指定する

```
~ 1 + a + b
~ -1 + a + b
```

ポイント

計画行列（design matrix）：model.matrixの出力

```
group      <- factor(c(rep("M", 3), rep("H", 3)))
replicates <- factor(c(1:3, 1:3))
model.matrix(~group+replicates)
```

```
  (Intercept) groupM replicates2 replicates3
1           1      1           0           0
2           1      1           1           0
3           1      1           0           1
4           1      0           0           0
5           1      0           1           0
6           1      0           0           1
attr(,"assign")
[1] 0 1 2 2
attr(,"contrasts")
attr(,"contrasts")$group
[1] "contr.treatment"
```

0と1の行列
contrasts

ポイント

計画行列（design matrix）：model.matrixの出力

$$O_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$


```
  (Intercept) groupM replicates2 replicates3
1           1      1           0           0
2           1      1           1           0
3           1      1           0           1
4           1      0           0           0
5           1      0           1           0
6           1      0           0           1
```

ポイント

線形モデルと計画行列の関係

$$O_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

i, jを書き下すと

$$O_{M,1} \sim \mu + \alpha_M + \beta_1 + \varepsilon_{M,1}$$
$$O_{M,2} \sim \mu + \alpha_M + \beta_2 + \varepsilon_{M,2}$$
$$O_{M,3} \sim \mu + \alpha_M + \beta_3 + \varepsilon_{M,3}$$
$$O_{H,1} \sim \mu + \alpha_H + \beta_1 + \varepsilon_{H,1}$$
$$O_{H,2} \sim \mu + \alpha_H + \beta_2 + \varepsilon_{H,2}$$
$$O_{H,3} \sim \mu + \alpha_H + \beta_3 + \varepsilon_{H,3}$$

ポイント

線形モデルと計画行列の関係

$$O_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

$$O_{M,1} \sim \mu + \alpha_M + \beta_1 + \varepsilon_{M,1}$$
$$O_{M,1} \sim \mu + \alpha_M + \alpha_H + \beta_1 + \beta_2 + \beta_3 + \varepsilon_{M,1}$$
$$O_{M,1} \sim \mu \times 1 + \alpha_M \times 1 + \alpha_H + \beta_1 + \beta_2 \times 0 + \beta_3 \times 0 + \varepsilon_{M,1}$$

contrasts: 1番目の水準の係数を０として残りと比較

$$O_{M,1} \sim \mu \times 1 + \alpha_M \times 1 + \beta_2 \times 0 + \beta_3 \times 0 + \varepsilon_{M,1}$$
$$1 \qquad 1 \qquad 0 \qquad 0$$

ポイント

計画行列, contrasts, 実験デザインの関係

観察数: 6

$$O_{M,1} \sim \mu + \alpha_M + \beta_1 + \varepsilon_{M,1}$$
$$O_{M,2} \sim \mu + \alpha_M + \beta_2 + \varepsilon_{M,2}$$
$$O_{M,3} \sim \mu + \alpha_M + \beta_3 + \varepsilon_{M,3}$$
$$O_{H,1} \sim \mu + \alpha_H + \beta_1 + \varepsilon_{H,1}$$
$$O_{H,2} \sim \mu + \alpha_H + \beta_2 + \varepsilon_{H,2}$$
$$O_{H,3} \sim \mu + \alpha_H + \beta_3 + \varepsilon_{H,3}$$

推定する係数の数: 6

$$\mu, \alpha_M, \alpha_H, \beta_1, \beta_2, \beta_3$$

推定したい係数の数よりも観察数が多くなくてはならない
contrasts: 1番目の水準の係数を0として残りと比較
→係数の数を削減

ポイント

計画行列, contrasts, 実験デザインの関係

観察数: 6

$$O_{M,1} \sim \mu + \varepsilon_{M,1}$$
$$O_{M,2} \sim \mu + \beta_2 + \varepsilon_{M,2}$$
$$O_{M,3} \sim \mu + \beta_3 + \varepsilon_{M,3}$$
$$O_{H,1} \sim \mu + \alpha_H + \varepsilon_{H,1}$$
$$O_{H,2} \sim \mu + \alpha_H + \beta_2 + \varepsilon_{H,2}$$
$$O_{H,3} \sim \mu + \alpha_H + \beta_3 + \varepsilon_{H,3}$$

推定する係数の数: **4**

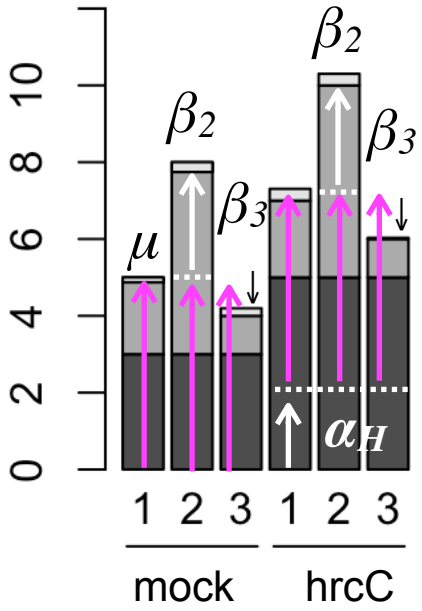
$$\mu, \quad \alpha_H, \quad \beta_2, \beta_3$$

推定したい係数の数よりも観察数が多くなくてはならない
contrasts: 1番目の水準の係数を0として残りと比較
→係数の数を削減

ポイント

線形モデルにおける係数推定のイメージ：最小二乗法を使いながら連立方程式を解く

$$
\begin{aligned}
O_{M,1} &\sim \mu + \varepsilon_{M,1} \\
O_{M,2} &\sim \mu + \beta_2 + \varepsilon_{M,2} \\
O_{M,3} &\sim \mu + \beta_3 + \varepsilon_{M,3} \\
O_{H,1} &\sim \mu + \alpha_H + \varepsilon_{H,1} \\
O_{H,2} &\sim \mu + \alpha_H + \beta_2 + \varepsilon_{H,2} \\
O_{H,3} &\sim \mu + \alpha_H + \beta_3 + \varepsilon_{H,3}
\end{aligned}
$$



$$\mu, \alpha_M, \alpha_H, \beta_1, \beta_2, \beta_3$$

ポイント

計画行列まとめ

$$O_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

```
  (Intercept) groupM replicates2 replicates3
1           1      1           0           0
2           1      1           1           0
3           1      1           0           1
4           1      0           0           0
5           1      0           1           0
6           1      0           0           1
```

ポイント

1. 0と1の意味
2. （この場合の）contrastsの概念: μ=replicate1の処理Hの係数
3. 観察数、実験デザインとの関連