dummy variable (indicator variable, coding)

- categorical (qualitative) predictors $\xleftrightarrow{\text{cf.}}$ categorical response (GLM)
  - nominal vs. ordinal ← e.g. educational attainment, disease diagnostic rating.
  - examples: male/female, treatment/control, eye colors, blocks, ...
  - qualitative in nature:
    - values are symbols, no quantitative meaning
    - no value exist between categories → relationship between y & $\underset{\sim}{x}$
  - **Q**: what properties can we explore for qualitative predictor?

    category $i \to y_{ij}$, $\mu_i = E(y_{ij}) \Rightarrow$ can only study *difference* between μ_i's ← "−"

    (cf., quantitative predictor) many categories (1-way, 2-way, ...) ⇒ many "types" of difference
  - **Q**: how to fit these predictors into the format of linear regression model $Y = X\beta + \varepsilon$? ⇒ Ans: dummy variables ← base functions for qualitative predictors
- one dichotomous predictor: two categories
  - for a dichotomous predictor C with two categories c_1 and c_2, define a dummy variable d:

    $d_c(x)$ in LNp 6 $\xleftrightarrow{\text{cf.}}$ $$d(C) = \begin{cases} 0, & \text{if } C = c_1, \\ 1, & \text{if } C = c_2. \end{cases}$$ ← this is a known function of C
  - for a data set with response y, one quantitative predictor x, and one qualitative predictor C (dummy variable d), possible models are: $\begin{matrix}1\\2\end{matrix} \rightrightarrows \begin{matrix}3\\4\end{matrix} \rightrightarrows 5$, "→" nested submodel $\omega \subset \Omega$

    model 1: $y = \beta_0 + \beta_1 d + \varepsilon$, model 2: $y = \beta_0 + \beta_1 x + \varepsilon$,

    model 3: $y = \beta_0 + \beta_1 d + \beta_2 x + \varepsilon$, model 4: $y = \beta_0 + \beta_1 x + \beta_2 xd + \varepsilon$,

    model 5: $y = \beta_0 + \beta_1 d + \beta_2 x + \beta_3 xd + \varepsilon$ ← interaction: a function of x & C

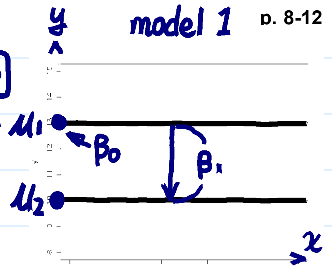
- **Q**: how to interpret β_i's in models 1~5?
  - model 1: $y = \beta_0 + \beta_1 d + \epsilon$ ← what difference?

    $C = c_1$: $\mu_1 = E(y|d=0) = \beta_0$

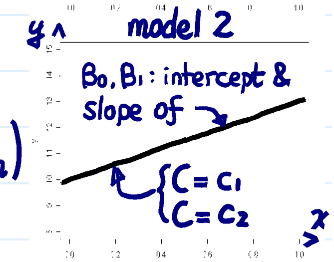
    $C = c_2$: $\mu_2 = E(y|d=1) = \beta_0 + \beta_1$

    $\Rightarrow$ $\beta_0 = \mu_1$, $\beta_1 = \mu_2 - \mu_1$

    reference



  - model 2: $y = \beta_0 + \beta_1 x + \epsilon$



  - model 3: $y = \beta_0 + \beta_1 d + \beta_2 x + \epsilon$ ← what difference? ← a main effect model (additive model, i.e., no interaction)

    cf. $(\beta_1 + \beta_2 d)x$ ← what difference?

    $C = c_1$: $\mu_{1,x} = E(y|d=0, x) = \boxed{\beta_0} + \beta_2 x$

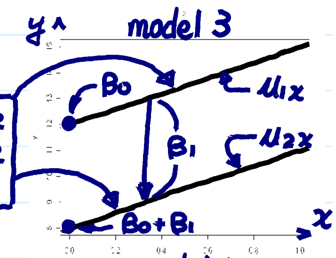
    $C = c_2$: $\mu_{2,x} = E(y|d=1, x) = \boxed{(\beta_0 + \beta_1)} + \beta_2 x$

    $\beta_0 = \mu_{1,0}$ (intercept in c_1 group)

    $\Rightarrow$ $\beta_1 = \mu_{2,x} - \mu_{1,x}$ (difference of intercepts)

    β_2 = slope (same slope in two categories)

    β2: slope of the 2 lines



  - model 4: $y = \beta_0 + \beta_1 x + \beta_2 (d \cdot x) + \epsilon$ ← an interaction

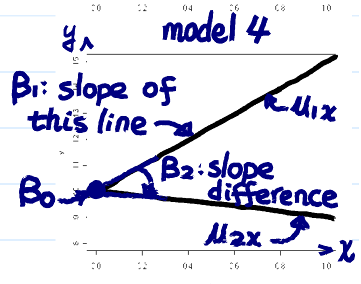
    $C = c_1$: $\mu_{1,x} = E(y|d=0, x) = \beta_0 + \boxed{\beta_1} x$ cf.

    $C = c_2$: $\mu_{2,x} = E(y|d=1, x) = \beta_0 + \boxed{(\beta_1 + \beta_2)} x$

    $\beta_0 = \mu_{1,0} = \mu_{2,0}$ (same intercept in two categories)

    $\Rightarrow$ β_1 = slope of category c_1

    β_2 = difference in slopes



- model 5: $y = \beta_0 + \beta_1 d + \beta_2 x + \beta_3 (d \cdot x) + \epsilon$ $(\beta_2+\beta_3 d)x$

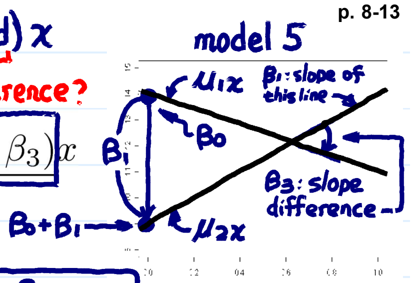
$C = c_1$: $\mu_{1,x} = E(y|d=0, x) = \boxed{\beta_0} + \boxed{\beta_2} x$ what difference?

$C = c_2$: $\mu_{2,x} = E(y|d=1, x) = \boxed{(\beta_0+\beta_1)} + \boxed{(\beta_2+\beta_3)} x$

reference line

$$\Rightarrow \begin{array}{rcl} \beta_0 & = & \mu_{1,0} \text{ (intercept of category } c_1) \\ \beta_2 & = & \text{slope of category } c_1 \\ \beta_1 & = & \text{difference in intercepts} \\ \beta_3 & = & \text{difference in slopes} \end{array}$$

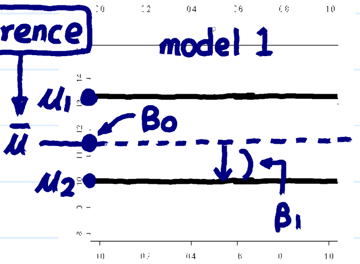
model 5



reference

e.g. with constant term

model 1



➢ alternative coding of dummy variable (better orthogonality)

location & scale change

$-2 \times \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} +1 \\ -1 \end{bmatrix} \to d(C) = \begin{cases} -1, & \text{if } C = c_1, \\ 1, & \text{if } C = c_2. \end{cases}$

Q: how to interpret β_i's in models 1~5 under this coding?

e.g. $\beta_3(d \times x)$ what difference?

- model 1: $y = \beta_0 + \beta_1 d + \epsilon$ **Q**: how about models 2~5? (exercise)

what difference?

$C = c_1$: $\mu_1 = E(y|d=-1) = \boxed{\beta_0} - \boxed{\beta_1}$

$C = c_2$: $\mu_2 = E(y|d=1) = \boxed{\beta_0} + \boxed{\beta_1}$

$$\Rightarrow \begin{array}{rcl} \beta_0 & = & (\mu_1+\mu_2)/2 \equiv \bar{\mu} \\ \beta_1 & = & (\mu_2-\mu_1)/2 = \mu_2 - \bar{\mu} \\ & & = -(\mu_1 - \bar{\mu}) \end{array}$$

➢ analysis strategy: start from the full model (model 5) if there are enough degrees of freedom, and then test if some terms can be eliminated

➢ identical methodology applies for more than 2 categories and more quantitative predictors

e.g. 2 quantitative predictors x_1, x_2

$E(y_x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$
$+ (\beta_0' + \beta_1' x_1 + \beta_2' x_2 + \beta_{11}' x_1^2 + \beta_{22}' x_2^2 + \beta_{12}' x_1 x_2) \times d$

➢ **Q**: what if data in the two categories have different variance?

$\boxed{\beta_{11} x_1^2 + \beta_{11}' d x_1^2 = (\beta_{11} + \beta_{11}' d) x_1^2}$

ANCOVA (共變數分析) $H_0 \cup H_A$ $\beta_1 = 0$: H_0

cf.

➢ *ANalysis of COVAriance*: testing model 3 (Ω) against model 2 (ω) (more than 2 categories and more quantitative predictors is allowed). The quantitative predictor is called *covariate* and is expected to have the same effect in all categories. The difference between categories is assumed to be an additive effect.

Ω: M3
ω: M1

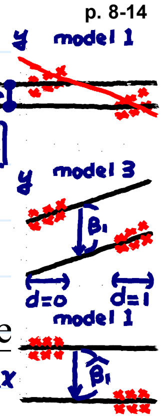
cf. blocking variable

$y = \beta_0 + \beta_1 d + \epsilon$

no interaction term btw x & d

cf. $y = \beta_0 + \beta_1 d + \beta_2 x + \epsilon$

$y - \hat{\beta}_2 x = \hat{\beta}_0 + \hat{\beta}_1 d + \hat{\epsilon}$



- one polytomous predictor: more than two categories

➢ for k categories, $k-1$ dummy variables are needed to depict the difference between categories (one parameter is used to represent constant term)

# of parameters = k

➢ various coding of dummy variables: 4 categories c_1, c_2, c_3, c_4 example

β-sample model (LNp.6-10)
$y = \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \beta_4 d_4 + \epsilon$

	d_1	d_2	d_3	d_4
c_1	1	0	0	0
c_2	0	1	0	0
c_3	0	0	1	0
c_4	0	0	0	1

$\beta_1 = \mu_1, \beta_2 = \mu_2$
$\beta_3 = \mu_3, \beta_4 = \mu_4$

(0,1) coding cf. treatment coding

dummy variables → base functions

known functions of C

reference

	d_1	d_2	d_3
c_1	0	0	0
c_2	1	0	0
c_3	0	1	0
c_4	0	0	1

Helmert coding (-1,1) coding

	d_1	d_2	d_3
c_1	−1	−1	−1
c_2	1	−1	−1
c_3	0	2	−1
c_4	0	0	3

sum coding

	d_1	d_2	d_3
c_1	−1	−1	−1
c_2	1	0	0
c_3	0	1	0
c_4	0	0	1

➢ consider the model: $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \epsilon$

- properties of treatment coding: what difference?

reference

$C = c_1$: $\mu_1 = E(y|d_1=0, d_2=0, d_3=0) = \beta_0$

$C = c_2$: $\mu_2 = E(y|d_1=1, d_2=0, d_3=0) = \beta_0 + \beta_1$

$C = c_3$: $\mu_3 = E(y|d_1=0, d_2=1, d_3=0) = \beta_0 + \beta_2$

$C = c_4$: $\mu_4 = E(y|d_1=0, d_2=0, d_3=1) = \beta_0 + \beta_3$

$$\Rightarrow \begin{array}{rcl} \beta_0 & = & \mu_1 \\ \beta_1 & = & \mu_2 - \boxed{\mu_1} \\ \beta_2 & = & \mu_3 - \boxed{\mu_1} \\ \beta_3 & = & \mu_4 - \boxed{\mu_1} \end{array}$$

- treats c_1 as a reference
- it is convenient if a "standard" categories exists
- d_1, d_2, and d_3 are mutually orthogonal, but not orthogonal to constant term

▪ properties of Helmert coding: $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \epsilon$

$$C = c_1: \quad \mu_1 = E(y|d_1=-1, d_2=-1, d_3=-1) = \beta_0 - \beta_1 - \beta_2 - \beta_3$$
$$C = c_2: \quad \mu_2 = E(y|d_1=1, d_2=-1, d_3=-1) = \beta_0 + \beta_1 - \beta_2 - \beta_3$$
$$C = c_3: \quad \mu_3 = E(y|d_1=0, d_2=2, d_3=-1) = \beta_0 + 2\beta_2 - \beta_3$$
$$C = c_4: \quad \mu_4 = E(y|d_1=0, d_2=0, d_3=3) = \beta_0 + 3\beta_3$$

What difference?

$$\Rightarrow \quad \beta_0 = \frac{\mu_1+\mu_2+\mu_3+\mu_4}{4} \equiv \bar{\mu}$$
$$\beta_1 = \frac{\mu_2 - \mu_1}{2}$$
$$\beta_2 = \frac{\mu_3 - ((\mu_1+\mu_2)/2)}{3}$$
$$\beta_3 = \frac{\mu_4 - ((\mu_1+\mu_2+\mu_3)/3)}{4}$$

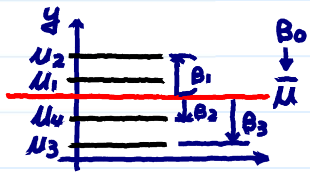
$$\underset{\sim}{\mu} = \begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\\ \mu_4\end{bmatrix} = \underbrace{\begin{bmatrix}1 & -1 & -1 & -1\\ 1 & 1 & -1 & -1\\ 1 & 0 & 2 & -1\\ 1 & 0 & 0 & 3\end{bmatrix}}_{A} \underbrace{\begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\end{bmatrix}}_{\beta}$$

$$\underset{\sim}{\beta} = A^{-1}\underset{\sim}{\mu}$$

- constant term, d_1, d_2, and d_3 are orthogonal when there are equal # of observations in each categories
- hard to interpret parameters
- may suitable for ordinal qualitative predictor

▪ properties of sum coding: $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \epsilon$

$$C = c_1: \quad \mu_1 = E(y|d_1=-1, d_2=-1, d_3=-1) = \beta_0 - \beta_1 - \beta_2 - \beta_3$$
$$C = c_2: \quad \mu_2 = E(y|d_1=1, d_2=0, d_3=0) = \beta_0 + \beta_1$$
$$C = c_3: \quad \mu_3 = E(y|d_1=0, d_2=1, d_3=0) = \beta_0 + \beta_2$$
$$C = c_4: \quad \mu_4 = E(y|d_1=0, d_2=0, d_3=1) = \beta_0 + \beta_3$$



What difference?

$$\Rightarrow \quad \beta_0 = \frac{\mu_1+\mu_2+\mu_3+\mu_4}{4} \equiv \bar{\mu}$$
$$\beta_1 = \mu_2 - \boxed{\bar{\mu}}$$
$$\beta_2 = \mu_3 - \boxed{\bar{\mu}} \leftarrow \text{reference}$$
$$\beta_3 = \mu_4 - \boxed{\bar{\mu}}$$

Q: Why no $\mu_1 - \bar{\mu}$?

$\beta_1+\beta_2+\beta_3 = \mu_2+\mu_3+\mu_4 - 3\bar{\mu}$
$= 4\bar{\mu} - \mu_1 - 3\bar{\mu} = -(\mu_1 - \bar{\mu})$
$\Rightarrow \mu_1 - \bar{\mu} = -(\beta_1+\beta_2+\beta_3)$

H_0: $\beta_1=\beta_2=\beta_3=0$
($\mu_1=\mu_2=\mu_3=\mu_4=\bar{\mu}$)
H_A: at least one of β_i's not 0
(at least one of μ_i's not $\bar{\mu}$)

- β_0 represent overall mean
- compare each category with the overall mean
- lesser orthogonal

$H_0(\omega)$: $y = \beta_0 + \varepsilon$, $H_0 \cup H_A(\Omega)$: $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \varepsilon$
↳ check codings in LNp.14

➢ Note: the choice of coding does not affect the R^2, $\hat{\sigma}$ and overall F-test (to test H_0: $\beta_1=\beta_2=\beta_3=0$, the three codings have same ω and Ω)

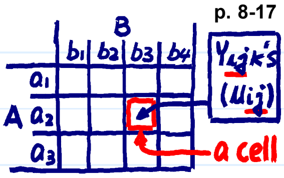
one qualitative predictor ← 變異數分析

➢ the overall F-test is one-way ANOVA (ANalysis Of VAriance) ← ANOVA does not depend on the choice of d_i's

➢ Q: how to work with quantitative predictors? ⇒ identical methodology as in 2 categories case. Q: how to interpret parameters in the case?

- two qualitative predictors → A & B can be crossing or nesting (cf.)

(say, A: I=3 categories a_1, a_2, a_3; B: J=4 categories, b_1, b_2, b_3, b_4)



➢ number of different category combinations = 3×4 =12 ← # of (A,B)=(a_i,b_j) combinations

denote their means as μ_{ij}, i=1, 2, 3 and j=1, 2, 3, 4, i.e.,

total df = $\sum_{i,j} n_{ij}$ but can only fit at most IJ=12 β's →
$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \quad k = 1, 2, \ldots, n_{ij},$$

	A	B	Y	
1	a_1	b_1	Y_{111}	μ_{11}
	⋮		⋮	
	a_1	b_1	$Y_{11n_{11}}$	
2	a_1	b_2	Y_{121}	μ_{12}
	a_1	b_2	$Y_{12n_{12}}$	
⋮			⋮	

n_{ij} = number of observations in category $A=a_i$ and $B=b_j$ → called replicates

➢ **Q**: how to depict the difference between μ_{ij}'s? → use dummy variables for A & B: A: d_1^A, d_2^A; B: d_1^B, d_2^B, d_3^B

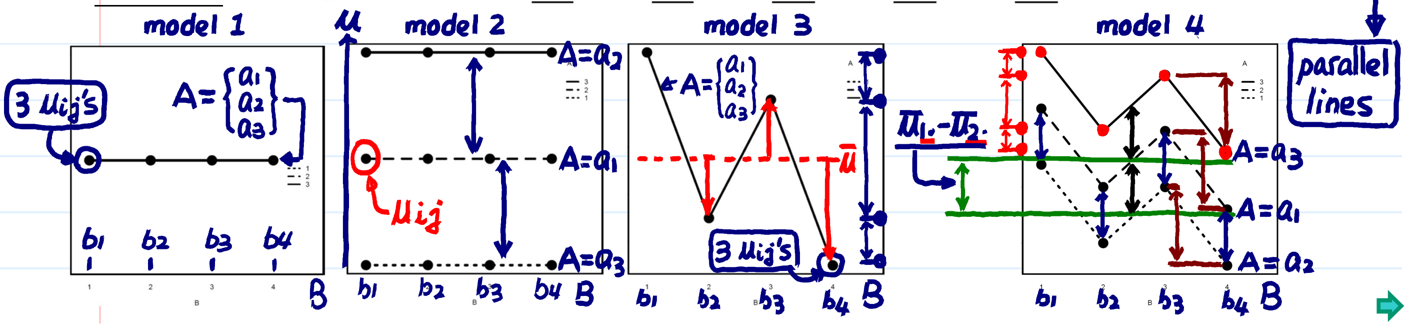
consider the following linear models:

- model 1: $E(y_{ijk}) = \beta_0$
- model 2: $E(y_{ijk}) = \beta_0 + \beta_1 d_1^A + \beta_2 d_2^A$
- model 3: $E(y_{ijk}) = \beta_0 + \beta_1 d_1^B + \beta_2 d_2^B + \beta_3 d_3^B$
- model 4: $E(y_{ijk}) = \beta_0 + \beta_1 d_1^A + \beta_2 d_2^A + \beta_3 d_1^B + \beta_4 d_2^B + \beta_5 d_3^B$ ← main-effect model (an additive model)

Q: What difference do their β's depict?

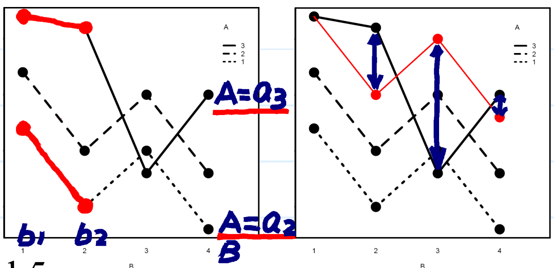
eg. $\mu_{11}-\mu_{21}=\mu_{12}-\mu_{22}=\mu_{13}-\mu_{23}=\mu_{14}-\mu_{24}$

$\mu_{13}-\mu_{14}=\mu_{23}-\mu_{24}=\mu_{33}-\mu_{34}$

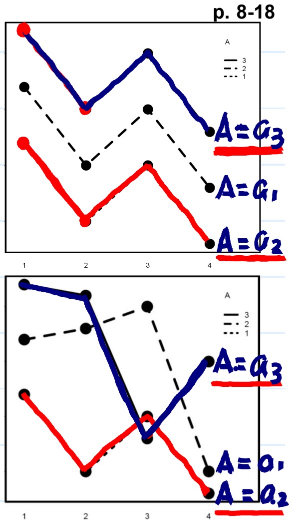


Q: how to perform interaction coding? what is interaction?

1 → 2, 3 → 4 → 5



add 3 (J-1) parameters and do it 2 (I-1) times ⇒ total (I-1)(J-1) parameters ← $\boxed{IJ-1-(I-1)-(J-1)}$



- model 5:

$$E(y_{ijk}) = \beta_0 + \beta_1 d_1^A + \beta_2 d_2^A + \beta_3 d_1^B + \beta_4 d_2^B + \beta_5 d_3^B$$
$$+ \sum_{i=1}^{2} \sum_{j=1}^{3} \beta_{ij} \boxed{d_{ij}} \longrightarrow \boxed{\text{2-factor interaction}}$$

Q: how to define d_{ij}'s? ① $d_{ij} = d_i^A \times d_j^B$ ② nested effect approach

$\boxed{\text{\# of parameters: } 1 + 2 + 3 + 6 = 12}$

interaction plot: replace μ_{ij}'s by cell means $\bar{y}_{ij} = \sum_{k=1}^{n_{ij}} y_{ijk}/n_{ij}$

effects enter the model "sequentially"

➢ Two-way (sequential) ANOVA (Type I) cf. t-test, F-test] discussed in LNp.4-10~16 (Type III)

(two qualitative predictors)

- anova(y~1+A+B+A:B)

test a batch of effects → H_0: A $\overset{?}{=}$ 0, H_0: B $\overset{?}{=}$ 0, H_0: A:B $\overset{?}{=}$ 0

1) test ω_1:model 1 (y~1) against Ω_1:model 2 (y~1+A) [$df_\omega - df_\Omega$=2] ← Type I

cf. ω: y~1+B+A:B, Ω: y~1+A+B+A:B ← Type III

$\Omega_k = \omega_{k+1}$

2) test ω_2:model 2 (y~1+A) against Ω_2:model 4 (y~1+A+B) [$df_\omega - df_\Omega$=3] ← Type I

cf. ω: y~1+A+A:B, Ω: y~1+A+B+A:B ← Type III

3) test ω_3:model 4 (y~1+A+B) against Ω_3:model 5 (y~1+A+B+A:B) [$df_\omega - df_\Omega$=6]

general F form in LNp.4-10 cf.

$$F=\frac{(RSS_\omega-RSS_\Omega)/(df_\omega-df_\Omega)}{RSS_{\text{model 5}}/df_{\text{model 5}}}\sim F_{df_\omega-df_\Omega,\,df_{\text{model 5}}}$$

$\hat\sigma^2_{\text{model 5}}=\hat\sigma^2_{\text{pure error}}$

- invariant to the choice of dummy variables if they generate same ω and Ω

- ANOVA could have different results when the order of effect sequence is changed, e.g., anova(y~1+B+A+A:B): cf. Type III invariant

$RSS_\omega-RSS_\Omega$ from $\Omega_{\text{model 5}}$, $RSS_{\text{model 5}}$ from $\Omega^\perp_{\text{model 5}}$ → independent

2) cf. α) test ω_1:model 1 (y~1) against Ω_1:model 3 (y~1+B) [$df_\omega-df_\Omega=3$]

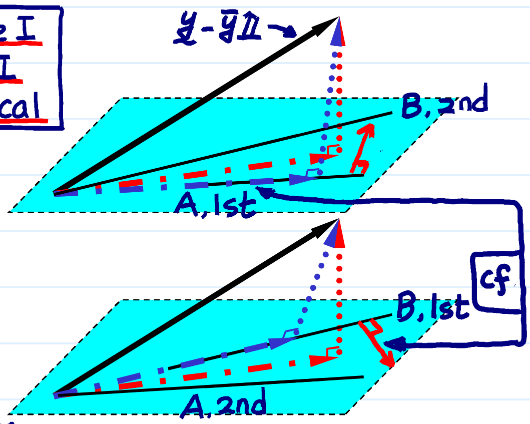
1) cf. β) test ω_2:model 3 (y~1+B) against Ω_2:model 4 (y~1+B+A) [$df_\omega-df_\Omega=2$]

χ) test ω_3:model 4 (y~1+B+A) against Ω_3:model 5 (y~1+B+A+A:B) [$df_\omega-df_\Omega=6$]

- anova(y~1+A+B+A:B) and anova(y~1+B+A+A:B) will have **identical** results when orthogonality exists between the 3 groups of effects: span{d_i^A's}, span{d_j^B's}, span{d_{ij}'s}, because in the case, $RSS_\omega-RSS_\Omega$ would equal for 1) and β), 2) and α), 3) and χ)

Also, Type I & Type III are identical

(I-1)(J-1)(K-1) parameters



- identical methodology applies for more qualitative (3-factor interaction, 4-factor interaction, …) and quantitative predictors (similar modeling to what in LNp.8-12~13)

❖ **Reading**: Faraway(2015, 1st ed.), chapter 13, 14.1, 15.1 ❖ **Further reading**: D&S, chapters 14, 23

Transformation

- Recall: check LNp.7-9~16. transformations for different purposes
  - objective: for some data, data after transformation can better fit a linear model (approximate)
  - **Q**: how to choose an appropriate transformation? → various plots ← rather subjective; cf → numerical method
  - transformation can be applied on response and on predictors
- transformation of response
  - Box-Cox transformation family: location & scale changes of y^λ; needed in likelihood approach

$$t_\lambda(y)=\begin{cases}(y^\lambda-1)/\lambda, & \text{if } \lambda\neq 0,\\ \log(y), & \text{if } \lambda=0.\end{cases}$$

collect a lot of transformations using a parameter λ; use data to determine

  - $t_\lambda(y)$ is continuous in λ: for fixed y>0, (also, differentiable)

$$\lim_{\lambda\to 0}t_\lambda(y)=\lim_{\lambda\to 0}(y^\lambda-1)/\lambda=\lim_{\lambda\to 0}(y^\lambda\log(y))/1=\log(y)$$

$\frac{dy_\lambda}{dy}=y^{\lambda-1}$

  - λ=1 ⇒ no transformation, λ=0 ⇒ log, $\lambda\neq$0 or 1 ⇒ power transformation
  - model: $Y_\lambda\equiv t_\lambda(y)=X\beta+\varepsilon$, $\varepsilon\sim N(\mathbf{0},\sigma^2 I)$ ⇒ $Y_\lambda=t_\lambda(Y)\sim N(X\beta,\sigma^2 I)$,
    - parameters: λ, β, σ likelihood: $\mathcal{L}(\lambda,\beta,\sigma;Y)=\mathcal{L}(\lambda,\beta,\sigma;Y_\lambda)\cdot|J|$
    - can write down likelihood for estimation and testing of λ
    - choice of transformation becomes a estimation/test problem — numerical method ⇒ more objective

profile log-likelihood

  - the log-likelihood is (exercise, check LNp.4-9; $\propto\hat\sigma^{-n}$) evaluate how far away from no transformation

$$L(\lambda)=\max_{\beta,\sigma^2}L(\lambda,\beta,\sigma|Y)\to L(\lambda)=\underbrace{(-n/2)\log(RSS_\lambda/n)}_{\text{goodness of fit}}+\underbrace{(\lambda-1)\,\Sigma\log(y_i)}_{\text{adjustment}}$$

β: $(X^TX)^{-1}X^TY_\lambda$, σ: $\sqrt{RSS_\lambda/n}$

Q: Why add this adjustment?

where RSS_λ = residual sum of square when using $t_\lambda(y)$ as response, i.e.,

$$RSS_\lambda=[t_\lambda(y)]^T(I-H)t_\lambda(y)=Y_\lambda^T(I-H)Y_\lambda$$