Week 6

categorical

ANOVA : Y ← X

salary ...... race gender

toothlength

day

1) Y continuous
2) Normality
3) Variance
4) Indep

Given

2-way ANOVA with interaction

[Toothlength] ~ (Dose + Supp + Dose * Supp)

→ 0.5 / 1 / 2

$SS_{Total} = [SS_{model} + SS_{res \cdot Error}$

$= [SS_{Dose} + SS_{supp} + SS_{D*S}] + SS_{Err}$

Dose effect

1) significance test [Ho: no dose effect ; Ha: dose effect]

↑ $F\text{-stat} = \dfrac{SS_{Dose} / df()}{SS_{Err} / df()}$

→ supplement effect

[Ho: no supplement effect ; Ha: supplement effect

$F\text{-stat} = \dfrac{SS_{supp} / df()}{SS_{Err} / df}$

Dose * Sup

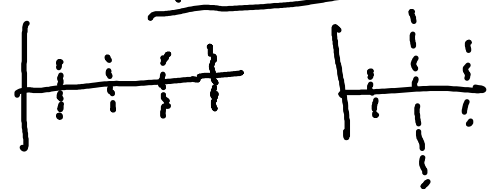
[Ho: no interaction between Dose-Supp ; Ha: interaction effect

$F\text{-stat} = \dfrac{SS_{D*S} /}{SS_{Err} /}$

2) R^2 $\left[\dfrac{SS_{model}}{SS_{Total}}\right]$

3) post-hoc test
pairwise-two-sample t-test

→ (④) Diagnostics — Normality
— Equal variance → Levene's test
→ H_0: —
H_1: —

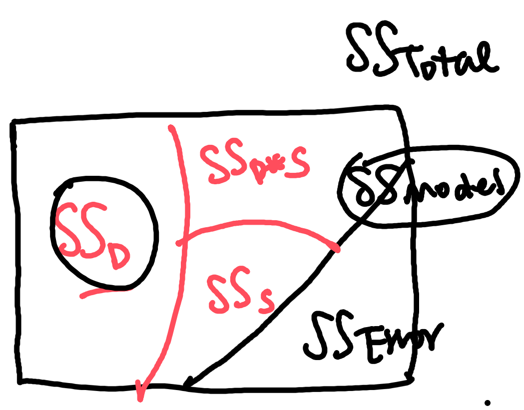


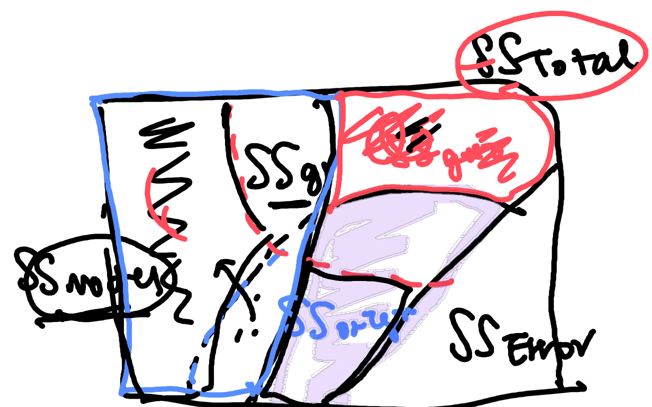
Balanced data

- terms are orthogonal
- SS model - unique decomposition

Unbalanced data

- terms are NOT orthogonal
- SS model





→ [Toothlength ~ D + S + D*S]

[days ~ grade + origin + g*o]

Type I — SS_D, SS_S, SS_{D*S} → SS_{model}

SS(grade)

SS(o|g) = SS(o)

SS(o*g|o.g) = SS(o*g)

Type I : sequential decomposition

	SS
grade	SS(grade) +
origin	SS(origin|grade) +
grade*origin	SS(g*o|g.o)

" SS_{model}

Type II

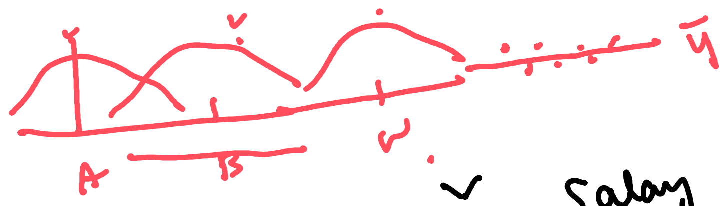
Type III

grade $SS(\text{grade} \mid o.\text{geo})$

origin $SS(o \mid g.\text{geo})$

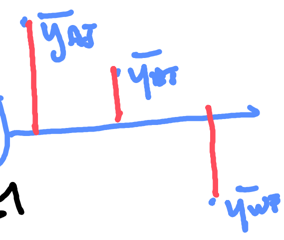
grade*origin $SS(\text{geo} \mid g.o)$

SSmodel



salary ~ race + gender

NOTE: orthogonal unorthogonal



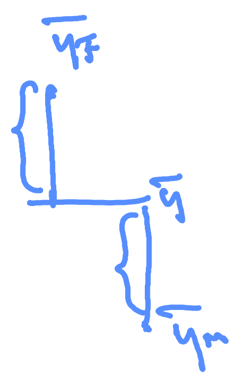
	A	B	W	
F	10	10	10	$\bar{y}_F$
M	10	10	10	$\bar{y}_m$
	$\bar{y}_A$	$\bar{y}_B$	$\bar{y}_W$	

$\bar{y}$

→ SS model $\bar{y}_{AF}$ $\bar{y}_{BF}$ - - -

SS race

SS gender



$\bar{y}_A$

	A	B	W
F	10	10	10
M	10	10	1
	$\bar{y}_A$	$\bar{y}_B$	$\bar{y}_W$

W Race

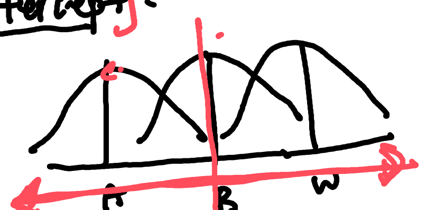
F

[SS intercept]? (from type III result)

$M_{tooth} = \boxed{} + D + S$

M_0 D_0 S

SS Total R^2



$R^2 \quad \frac{SS_{model}}{SS_{Total}}$ [type I type II] post-hoc test interpretation examples

A B C D

1. $Y \sim \boxed{X}$. $D > B$

2: $Y \sim X$ $[M_1 \quad M_2 \quad M_3 \quad \boxed{M_4}]$

${}_4C_2$

✓ $M_1 \neq M_4$

✓ $M_2 \neq M_4$

$M_3 \neq M_4$

M_4

3. $Y \sim X$ Fed 1 F2 F3 F4

$F_2 > F_1$

$F_3 > F_1$

$F_4 > F_1$

$F_3 > F_2$

$F_4 > F_2$

$F_4 < F_3$

$F_3 > F_4 > F_2 > F_1$

4.

Model Selection

50 50

Y ← [x_1 x_2 x_3 x_4]

Salary, gender, race, region

(i) Backward Elimination

- Eliminate ~~to~~ the least significant variables
  large pvalues
- Stopping rule: p-value 0.05

- Full model

$Y \sim x_1 + x_2 + x_3 + x_4$

	SS - F	p-val
x_1		0.03
x_2		0.78
x_3		0.58
x_4		0.01

→ remove x_2

x_2

$Y \sim x_1 + x_3 + x_4$

	p-val
x_1	0.01
x_3	0.27
x_4	0.03

remove x_3

x_3

0.1

$Y \sim x_1 + x_4$

	p-val
x_1	0.001
x_4	0.03

0.06

0.06 p-val < 0.05

p-value 0.05

Forward selection
→ Add most significant variable (smallest p-val)
→ stop (p-va 0.05)

$Y \sim X_1$ | $Y \sim X_2$ | $Y \sim X_3$ | $Y \sim X_4$

X_1	0.004 ✓

X_2	0.89

X_3	0.75

X_4	0.01

add X_1

X_1 $\underbrace{X_2 \quad X_3 \quad X_4}$

→ $Y \sim X_1 + X_2$ | $Y \sim X_1 + X_3$ | $Y \sim X_1 + X_4$

X_1	0.01 (0.5) ×
X_2	0.27

X_1	0.03 ×
X_3	0.58

× X_1	0.12
X_4	0.03

add X_4

X_2. X_3

→ $Y \sim \underbrace{X_1 + X_4} + X_2$

X_1	0.01 ✓ ✓
X_2	0.47
X_4	~~0.03~~ ✓

0.07

$Y \sim \underbrace{X_1 + X_4} + X_3$

$\boxed{\cancel{X_3}}$

X_1	0.009 ✓
× ✓ X_3	0.38 ~~0.58~~
X_4	0.02 ✓

$Y \sim X_1 + X_4$

p   r AIC

Model Selection (FS / BE)

- One by one
- criteria (stopping rule) pval 0.05

  x_3 0.03

  stopping rule pval 0.01 — BE : smaller, FS : smaller

  " 0.1 — BE larger, FS :

  0.07

  (note) large model # many # var

  small model # small # var

- Final models from FS / BE can be different

- Interaction

  4-way

  4

  $_4C_2$ → 15

Model Selection (FS/BE) only with main effects

x_1 x_2 x_3 x_4

↓

Determine Final main effect model

$Y \sim x_1 + x_4$

↓

check interaction

$Y \sim x_1 + x_4$

$Y \sim x_1 + x_4 + x_1 x_4$

when interaction is significant, but main effects are not.

10/5 6pm

$\boxed{\text{~~10/15 6pm~~}}$

Midterm

Quiz

programming