Analysis of variance (ANOVA)

TMA4267 L15
10.03.2017

Ex: Concrete recipes : 5 recipes to produce concrete, tested on 6 samples each. Measured moisture. Q: is there a difference between the recipes wrt moisture.

$\Updownarrow$

is the variability between the recipes large compared variability within.

1) One-way ANOVA model

$$Y_{ij} = \mu_i + \varepsilon_{ij} \qquad \begin{array}{l} i = 1, \dots, p \\ j = 1, \dots, n_i \end{array} \qquad \begin{array}{l} \text{Ex: } p = 5 \\ n_i = 6 \ \forall i \end{array}$$

$\varepsilon_{ij} \sim N(0, \sigma^2)$ and ε_{ij}'s independent

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

μ: grand mean

$$\mu_i = \mu + \alpha_i \Leftrightarrow \alpha_i = \mu_i - \mu$$

α_i: difference to grand mean

Q: how can we write this as a linear regression?
$(Y = X\beta + \varepsilon)$

$$n = \sum_{i=1}^{p} n_i, \quad \text{Ex: } 5 \cdot 6 = 30$$

Yes, the model can be fitted as a linear regression with parameters $(\underset{\beta_0}{\mu}, \underset{\beta_1}{\alpha_1}, \underset{\beta_2}{\alpha_2}, \dots, \alpha_p)$.

Previously: dummy variable coding. ← problem: we want $\hat{\mu}$ average of all measurements

Now: effect coding

Impose a restriction on the α's : sum-to-zero-constraint $\sum_{i=1}^{p} \alpha_i = 0$, in practice only use $\alpha_1, \dots, \alpha_{p-1}$ and let $\alpha_p = -\sum_{i=1}^{p-1} \alpha_i$. This gives effect-coding of design matrix

Ex:

$$\underset{30\times 5}{X} = \begin{array}{c} \beta_0=\mu \quad \beta_1=\alpha_1 \quad \beta_2=\alpha_2 \quad \beta_3=\alpha_3 \quad \beta_4=\alpha_4 \quad \left[\alpha_5 = -(\alpha_1 + \dots + \alpha_4)\right] \\ \left[\begin{array}{ccccc} 1 & 1 & 0 & 0 & 0 \\ & & & & \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 & -1 \end{array}\right] \begin{array}{l} \text{recipe 1} \\ \text{6 identical rows} \\ \text{recipe 2} \\ \text{recipe 3} \\ \text{recipe 4} \\ \text{recipe 5} \end{array} \end{array}$$

Then, use the "old" $Y = X\beta + \varepsilon$, $\varepsilon \sim N_n(0, \sigma^2 I)$

and $\beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_{p-1} \end{bmatrix}$, and $\hat{\beta} = (X^T X)^{-1} X^T Y$.

$$\hat{\alpha}_p = -(\hat{\alpha}_1 + \dots + \hat{\alpha}_{p-1})$$

2) Hypothesis test

$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_p$ vs H_1: at least one different

$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_p = 0$ vs H_1: at least one $\neq 0$.

Q: How can we do this with linear hypotheses and F_{obs} from L14?

Solution a: Write as linear hypotheses:

$\mu, \alpha_1 \dots, \alpha_{p-1}$

number of parameters: p

"$C\beta = d$" $\quad \beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_4 \end{bmatrix}$

Ex:

$$\underset{\substack{r \times p \\ \shortparallel \\ 4 \times 5}}{C} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad d = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$n - p = 30 - 5$

$f_{obs} = 4.3$, $\quad$ p-value $= P(F_{4,25} > 4.3) = 0.00875$

$\Rightarrow$ reject H_0 $\Rightarrow$ difference between recipes.

Solution b) : in R, fit full model and
anova(fullmodel). ANOVA table

"SS"	df	SS	$\frac{SS}{df} = MS$	Fval	p-val
(SSR) Treatment (regression)	$r = p-1$	x	x	x	x
(SSE) Error	$n-p$	x	x		

<u>The classical way</u>

$Y_{..} =$ $\overset{\text{total}}{\text{average}}$ of data

$Y_{i.} =$ average grp i

SSA = our SSRegression, SSE = as before

$\Rightarrow$ F-test.

Two factor experiments

Ex: machine α's machines

γ's operator

$$Y_{ij} = \mu + \alpha_i + \gamma_j + \varepsilon_{ij}$$

$i = 1, .., r^*$ $= n_i \; \forall i$
$j = 1, .., s$
same (not n_i as before)

$\varepsilon_{ij} \sim N(0, \sigma^2)$ independent

We use sum-zero-constraint both for α's and γ's

H_0^A: $\alpha_1 = \alpha_2 = \dots = \alpha_r = 0$ effect of machine

Ex: machine: $r^* = 4$, $n = 4 \cdot 6 = 24$, #param: $1 + 3 + 5 = 9$

$$\underset{3\times 9}{C_A} = \begin{bmatrix} 0 & 1 & 0 & 0 & \\ 0 & 0 & 1 & 0 & \underset{3\times 5}{0} \\ 0 & 0 & 0 & 1 & \end{bmatrix} \rightarrow \underset{\substack{\uparrow \\ 3,15}}{F_A} = 3.34, \ p\text{-value} = 0.048$$

H_0^B: $\gamma_1 = \gamma_2 = \dots = \gamma_s = 0$ effect of operator

Ex: $s = 6$

$$\underset{5\times 9}{C_B} = \begin{bmatrix} 0 & & 1 & 0 & 0 & 0 & 0 \\ 0 & & 0 & 1 & 0 & 0 & 0 \\ 0 & \underset{5\times 3}{0} & 0 & 0 & 1 & 0 & 0 \\ 0 & & 0 & 0 & 0 & 1 & 0 \\ 0 & & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \underset{\substack{\uparrow \\ 5,15}}{F_B} = 5.29, \ p\text{-value} = 0.005$$

We have used "$C\beta = d$" for hypothesis test, but in practise you may use ANOVA ← already implemented in R:

```
fit = lm(y~A+B)
anova(fit)
```

Additive effects and interactions

$$Y_{ij} = \mu + \alpha_i + \gamma_j + \varepsilon_{ij} \quad \text{additive}$$

$$Y_{ijk} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \varepsilon_{ijk}$$

$i = 1, \dots, r$ (Ex: 2)
$j = 1, \dots, s$ (Ex: 5)
$k = 1, \dots, n_{ij}$ (Ex: 10)

$$\sum_{i=1}^{r} \alpha_i = 0$$

$$\sum_{j=1}^{s} \gamma_j = 0$$

$$\sum_{i=1}^{r} (\alpha\gamma)_{ij} = 0 \quad \text{for all } j$$

$$\sum_{j=1}^{s} (\alpha\gamma)_{ij} = 0 \quad i$$

In R: Words ~ Age * Process

Age + Process + Age : Process

cAge = −1 = young
+1 = old

First test: Interaction present? Ex: Yes!

Yes → compare combinations

No → Check effect of A: α_is
B: γ_js