XIII. Analysis of Variance

- only <u>categorical</u> covariates

$Z = (z_1, \dots, z_p)^T$, $z_j \in \mathcal{Z}_j$ = finite, low cardinality

---

<u>Main usage</u> <u>agricultural</u>

- designed (industrial) experiments
- controlled clinical trials

Z = p factors that may influence outcome

e.g. z_1 = type of production machine (A, B, C)

z_2 = production temperature (experiment done at prespecified temp. values of 20°C, 25°C, 30°C)

z_3 = used amount of some ingredient (few prespecified values used at experiment)

...

→ covariates are mostly <u>fixed</u>, not random, but this does not matter for theory (we condition by covariate values anyway)

13.1 One-way classification

≡ REPETITION from Mathematical Statistics 1 and from Chapter 4

$p=1$, $Z \in \mathcal{Z} = \{1, \dots, G\}$

$$E(Y|Z=g) = m(g) =: m_g \quad , g = 1, \dots, G$$

without loss of generality

→ data sorted by Z values

→ double subscript used

$$Z = \begin{pmatrix} z_1 \\ \vdots \\ z_{n_1} \\ \hline z_{n_1+1} \\ \vdots \\ z_{n_1+n_2} \\ \hline \vdots \\ \hline z_{n_1+n_2+\dots+n_{G-1}+1} \\ \vdots \\ \underbrace{z_{n_1+\dots+n_{G-1}+n_G}}_{n} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \hline 2 \\ \vdots \\ 2 \\ \hline \vdots \\ \hline G \\ \vdots \\ G \end{pmatrix}, \quad Y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1,n_1} \\ \hline y_{2,1} \\ \vdots \\ y_{2,n_2} \\ \hline \vdots \\ \hline y_{G,1} \\ \vdots \\ y_{G,n_G} \end{pmatrix} = \begin{pmatrix} y_1 \\ \hline y_2 \\ \hline \vdots \\ \hline y_G \end{pmatrix}$$

→ by linear model, we "model"

$$E(Y|Z) = \begin{pmatrix} m_1 \mathbb{1}_{n_1} \\ \vdots \\ m_G \mathbb{1}_{n_G} \end{pmatrix}$$

ASSUMPTION (if LM used):

$$\mathrm{var}(Y|Z) = \sigma^2 \cdot I_n$$

(homoscedasticity)

We will also assume

$$P(n_g > 0) = 1 \quad \forall g = 1, \dots, G$$

13.1.1 Parameters of interest

(1) Differences between the group means

$\theta_{g,h} := m_g - m_h \quad , g,h = 1,\dots,G, \; g \neq h$

m_g, m_h: elements of a vector $E(Y|Z)$

$\Rightarrow$ LSE of $\theta_{g,h}$, $\hat{\theta}_{g,h} = \hat{Y}_g - \hat{Y}_h$

$\hat{Y}_g, \hat{Y}_h$: fitted values for obs. with $Z=g$ and $Z=h$

$\rightarrow$ principal null hypothesis to be tested

$H_0: \; m_1 = \dots = m_G$

$\theta_{g,h} = 0 \quad , g,h = 1,\dots,G \; , \; g \neq h$

(2) Factor effects

Def 13.1 Factor effects in a one-way classification

By factor effects in case of a one-way classification we understand the quantities defined as

$\eta_g = m_g - \bar{m} \quad , g = 1,\dots,G$, where

$\bar{m} = \frac{1}{G} \sum_{h=1}^{G} m_h$ is the mean of the group means.

NOTE: $\bar{m} = \frac{1}{G} \sum_{h=1}^{G} E(Y|Z=h) = EY$ iff $P(Z=h) = \frac{1}{G} \; \forall h = 1,\dots,G$

$\bar{m}$ = mean result of experiment if each setting is used equally often

13.1.2 One-way ANOVA model

6

Regression space

$$\left\{ \begin{pmatrix} m_1 \mathbb{1}_{n_1} \\ \vdots \\ m_G \mathbb{1}_{n_G} \end{pmatrix} : m_1, \dots, m_G \in \mathbb{R} \right\} \subseteq \mathbb{R}^n$$

if $n_g > 0 \ \forall g \Rightarrow$ rec-dim $= G$

7

Full-rank parameterizations

$$m_g = \beta_0 + \mathbb{C}_g^T \beta^Z \quad , g = 1, \dots, G$$

$$\beta = (\beta_0, (\beta^Z)^T)^T$$

$$(\beta_1^Z, \dots, \beta_{G-1}^Z)^T$$

$$\mathbb{C} = \begin{pmatrix} \mathbb{C}_1^T \\ \vdots \\ \mathbb{C}_G^T \end{pmatrix} = \text{chosen (pseudo) contrast matrix with } G-1 \text{ columns}$$

NOTE: With $\mathbb{C} = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & & \\ 0 & \dots & 1 \\ -1 & \dots & -1 \end{pmatrix}$ = sum contrasts

$$\alpha_0 := \beta_0 = \frac{1}{G} \sum_{g=1}^{G} m_g = \bar{m}$$

factor effects:

$$\alpha_g := \beta_g = m_g - \bar{m} = \eta_g \quad , g = 1, \dots, G-1$$

$$\alpha_G := -\sum_{g=1}^{G-1} \beta_g = m_G - \bar{m} = \eta_G$$

13.1.3 Least squares estimation

Lemma 13.1 Least squares estimation in one-way ANOVA (4)

short: $\hat{m}_g = \hat{y}_{g,j} = \frac{1}{n_g}\sum_{\ell=1}^{n_g} y_{g\ell} =: \bar{y}_{g\cdot}$

$g = 1, \dots, G, \; j = 1, \dots, n_g.$

$$\hat{m} := \begin{pmatrix} \hat{m}_1 \\ \vdots \\ \hat{m}_G \end{pmatrix} = \begin{pmatrix} \bar{y}_{1\cdot} \\ \vdots \\ \bar{y}_{G\cdot} \end{pmatrix} \qquad \hat{y} = \begin{pmatrix} \bar{y}_{1\cdot} \mathbb{1}_{n_1} \\ \vdots \\ \bar{y}_{G\cdot} \mathbb{1}_{n_G} \end{pmatrix}.$$

Under normality

$$\hat{m} \mid Z \sim N_G(m, \sigma^2 V), \quad V = \begin{pmatrix} \frac{1}{n_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{n_G} \end{pmatrix}.$$

Proof: Possible full-rank parameterization is

$$E(y \mid Z) = \begin{pmatrix} m_1 \mathbb{1} \\ \vdots \\ m_G \mathbb{1} \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbb{1}_{n_1} & & 0 \\ & \ddots & \\ 0 & & \mathbb{1}_{n_G} \end{pmatrix}}_{X} \cdot \underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_G \end{pmatrix}}_{\beta}$$

i.e. $\beta_g = m_g, \; g = 1, \dots, G$

$$\Rightarrow (X^T X) = \begin{pmatrix} n_1 & & 0 \\ & \ddots & \\ 0 & & n_G \end{pmatrix}, \quad (X^T X)^{-1} = \begin{pmatrix} \frac{1}{n_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{n_G} \end{pmatrix}$$

$$X^T y = \begin{pmatrix} \sum_j y_{1j} \\ \vdots \\ \sum_j y_{Gj} \end{pmatrix} \Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y = \begin{pmatrix} \bar{y}_{1\cdot} \\ \vdots \\ \bar{y}_{G\cdot} \end{pmatrix}$$

$\hat{\beta} = \hat{m}$

$$\hat{y} = X\hat{\beta} = \begin{pmatrix} \bar{y}_{1\cdot} \\ \vdots \\ \bar{y}_{1\cdot} \\ \hline \vdots \\ \hline \bar{y}_{G\cdot} \\ \vdots \\ \bar{y}_{G\cdot} \end{pmatrix}$$

$$\text{var}(\hat{m} \mid \mathbb{Z}) = \text{var}(\hat{\beta} \mid \mathbb{Z}) = \sigma^2 (X^T X)^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{n_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{n_G} \end{pmatrix}$$

$\Downarrow$

normality $\equiv$ general LSE theory

LSE of regression coefficients and their linear combinations

Full-rank parameterization

$$m_g = \beta_0 + \mathbf{c}_g^T \boldsymbol{\beta}^z \quad \longleftrightarrow \quad \mathbf{m} = \beta_0 \mathbf{1} + \mathbf{C}\boldsymbol{\beta}^z$$

$$\begin{pmatrix} \bar{y}_{1\cdot} \\ \vdots \\ \bar{y}_{G\cdot} \end{pmatrix} = \hat{\mathbf{m}} = (\text{subvector of fitted values}) = \hat{\beta}_0 \mathbf{1} + \mathbf{C}\hat{\boldsymbol{\beta}}^z$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

$$\hat{\mathbf{m}} = \underbrace{(\mathbf{1}, \mathbf{C})}_{\text{invertible}} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}^z \end{pmatrix} \qquad \begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}^z \end{pmatrix} = (\mathbf{1}, \mathbf{C})^{-1} \hat{\mathbf{m}}$$

$\hat{\mathbf{m}}$: vector of group sample means

- in any parameterization
  - $\hat{\boldsymbol{\beta}}$ = linear combination of group sample means
  - the same holds for any linear combination of β's, e.g.

* $\theta_{gh} = m_g - m_h \quad \rightarrow \quad \hat{\theta}_{gh} = \hat{m}_g - \hat{m}_h = \bar{y}_{g\cdot} - \bar{y}_{h\cdot}$

* $\eta_g = m_g - \bar{m} \quad \rightarrow \quad \hat{\eta}_g = \hat{m}_g - \frac{1}{G}\sum_{h=1}^{G} \hat{m}_h =$

$$= \cancel{m_g}\ \bar{y}_{g\cdot} - \frac{1}{G}\sum_{h=1}^{G} \bar{y}_{h\cdot}$$

13.1.4 Within and between sums of squares, ANOVA F-test

Sums of squares

(usual) notation: $\bar{y} = \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} y_{g,j} = \frac{1}{n}\sum_{g=1}^{G} n_g \bar{y}_{g\cdot}$

Within groups sum of squares (= residual sum of sq.)

$$SS_e = \|y - \hat{y}\|^2 = \sum_{g=1}^{G}\sum_{j=1}^{n_g} (y_{g,j} - \hat{y}_{g,j})^2 = \sum_{g=1}^{G}\sum_{j=1}^{n_g} (y_{g,j} - \bar{y}_{g\cdot})^2$$

$$r_e = n - G$$

Between groups sum of squares (= regression sum of sq.)

$$SS_R = \|\hat{y} - \bar{y}\|_n\|^2 = \sum_{g=1}^{G}\sum_{j=1}^{n_g} (\hat{y}_{g,j} - \bar{y})^2 = \sum_{g=1}^{G} n_g (\bar{y}_{g\cdot} - \bar{y})^2$$

$$r_R = G - 1$$

One-way ANOVA F-test

consider $M: y|X \sim \mathcal{N}(X\beta, \sigma^2 I_n)$ ← arbitrary parameterization of one-way ANOVA

$M_0: \quad \sim \mathcal{N}(\mathbb{1}\beta_0, \sigma^2 I_n)$

$M_0 \subset M$

$$SS_e^0 = SS_T = \sum_{g=1}^{G}\sum_{j=1}^{n_g} (y_{g,j} - \bar{y})^2$$

$$\underbrace{SS_T}_{SS_e^0} = SS_e + SS_R$$

$$\Rightarrow SS_R = SS_e^0 - SS_e$$

$\Rightarrow$ F-statistic to test $M_0 \subset M$:

$$F = \frac{\left(\frac{SS_R}{G-1}\right)}{\left(\frac{SS_e}{n-G}\right)}$$

MS_R: between variability

MS_e: within variability

= classical ANOVA F-statistic

→ One-way ANOVA table

Effect (Term)	Degrees of freedom	Effect sum of squares	Effect mean square	F-stat.	P-value
Factor	$G-1$	SS_R (between)	MS_R	F	p
Residual	$n-G$	SS_e (within)	MS_e		